Robust Speech Recognition via Adaptation for German Oral History Interviews

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Michael Gref
aus
Kistakuz, Tadschikistan

Bonn, 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. rer. nat. Sven Behnke
2. Gutachter:	apl. Prof. Dr. rer. nat. Frank Kurth
Tag der Promotion:	30.09.2022
Erscheinungsjahr:	2022

Abstract

Automatic speech recognition systems often achieve remarkable performance when trained on thousands of hours of manually annotated and time-aligned speech. However, when applied in other conditions and domains than they were trained on, the systems' recognition quality often deteriorates, substantially limiting their real-world application. One of these applications is the automatic transcription of oral history interviews, i.e., interviews with witnesses of historical events. For the past twenty years, oral history interviews have been among the most challenging use cases for speech recognition due to a lack of representative training data, diverse and often poor recording conditions, and the spontaneous and occasionally colloquial nature of the speech.

This thesis proposes and studies the combination of different domain adaptation approaches to overcome the lack of representative training data and cope with the unpredictability of oral history interviews. We employ and investigate data augmentation to adapt broadcast training data to cover the challenging recording conditions of oral history interviews. We compare data augmentation approaches to conventional speech enhancement. To improve the system's performance further, we study domain adaptation via fine-tuning to adapt the acoustic models trained robustly on thousands of hours of annotated speech using a minimal amount of manually transcribed oral history interviews. We employ automatic transcript-alignment to generate adaptation data from transcribed but not time-aligned interviews and investigate the influence of different adaptation data sizes on domain overfitting and generalization. We reduce domain overfitting and improve the generalization of the adapted models employing cross-lingual adaptation in a multi-staged setup to leverage the vast availability of English speech corpora. Additionally, in this thesis, a human word error rate for German oral history interviews recorded under clean conditions is experimentally estimated to study and highlight the challenges of transcription even for humans and put current results of automatic transcription into perspective.

The proposed methods are evaluated on a representative oral history test set for the target domain and several additional German test sets from different domains. With this evaluation, we assure high robustness, obtain a reliable estimate of the real-world performance for conditions not seen in training, and avoid selecting models that suffer from domain overfitting. Overall, we halved the word error rate

compared to the baseline using the proposed methods, simultaneously improving the recognition performance on the other domains by a substantial margin.

Acknowledgements

This thesis was written in collaboration with the NetMedia Department of the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), the University of Bonn, and the Faculty of Electrical Engineering and Computer Science of the Niederrhein University of Applied Sciences.

First, I would like to thank Prof. Sven Behnke for kindly accepting me as his doctoral student at the University of Bonn. His trust in the concept made this thesis possible in the first place. I thank him for his interest in my research and encouragement regarding the thesis topic. His immense knowledge, guidance, and insightful discussions have played a key role in my work.

Furthermore, I would like to express my sincere gratitude to my advisor Dr. Christoph Schmidt at the Fraunhofer IAIS for his continuous support, insightful discussions, motivation, and immense knowledge. Thank you for always taking the time to do this despite your numerous commitments.

I would also like to express my special thanks to Prof. Hans-Günter Hirsch of the Niederrhein University of Applied Sciences and Dr. Joachim Köhler of the Fraunhofer IAIS. This collaboration, my research, and, eventually, this thesis would never have come about in this form without both of you believing in me and giving me this chance. Thank you for this and for encouraging me to do things that I would not have thought myself capable of doing five years ago that developed me as a researcher, educator, and human being.

Sincere thanks go to Prof. Roger Frese of the Düsseldorf University of Applied Sciences. In 2014, you paved the way for my academic career by hiring me as your research assistant and encouraging me to pursue a Ph.D., even though circumstances still pointed to a very uncertain future. This time as your research assistant also sparked my love for academic teaching and digital signal processing, which persists to this day.

Furthermore, I would like to thank all my former and current colleagues and friends I had the pleasure to meet and work with along the way at the Fraunhofer IAIS, Niederrhein and Düsseldorf Universities of Applied Sciences. Very special thanks go to Prof. Steffen Goebbels for his tireless willingness to discuss even the most arbitrary mathematical topics with me, even if they had nothing to do with my actual research work. I also thank Steffen for one of the biggest flexes so far: appearing first in the acknowledgments of a mathematics textbook. I want to thank Dr. Michael Stadtschnitzer for impressively proving that finishing a Ph.D. at

Fraunhofer IAIS is actually possible. I would also like to thank my other long-time colleagues at Fraunhofer, Dr. Oliver Walter, Dr. Paul Wallbott, Dr. Jan-Gerrit Richter, David Laqua, Jens Fröschel, Julia Pritzen, Falk Jaeger, and Carolina Held, for regularly asking me about the current status of my thesis, providing distractions when necessary, and for their encouragement. Further thanks go to Alexander Bathe and Stefan Lörcks for the unforgettable and formative journey into the wonderland of Abelian groups, polynomials over finite fields, and holomorphic functions. Actually, this made my work more difficult than straightforward, but I would not change a day. May our paths soon meet again where they parted.

Unfortunately, I cannot list every single person I have met on my journey so far. Therefore, particularly to you, who are reading this: Thank you.

Large parts of this work were carried out as part of the joint project KA3.[1] I thank everyone involved in this exciting research project and the great collaboration. In particular, I would like to thank Dr. Almut Leh from the Archive *Deutsches Gedächtnis*, University of Hagen, for providing the exciting, challenging use case for my research, the valuable discussions, and constructive feedback. This work has given me the feeling that my work is not only of theoretical interest also but contributes to making other people's research easier.

Furthermore, I would like to thank Dr. Ruth Rosenberger, Nike Matthiesen, and Jonathan Heil from the Haus der Geschichte Foundation for their great cooperation and support in our joint research project[2], which enabled me to finalize my research work and explore new research directions.

I thank my family, Ida and Nikolaus Gref, my sisters Eugenia and Irene Gref, as well as my friends for the motivating words and the necessary balance during my doctoral studies. Thanks and a wholehearted *"Moin"* to my wife's family in East Frisia for hosting and organizing our wedding last year and giving us one of the best days of our lives during a very stressful time in our doctorate work. Thank you!

My deepest gratitude goes to my wife, Dr. Imke Busboom, for her wholehearted support and unconditional love. Who would have thought ten years ago the two of us would be here today at this point? Here's to the next chapter.

[1] This research project and parts of this thesis have 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 numbers 01UG1511B and 01UG1811B.

[2] This research project and parts of this thesis have been funded by the German Federal Government Commissioner for Culture and Media.

Contents

1 Introduction

Nowadays, automatic speech recognition (ASR) is probably one of the best-known and most successful disciplines of artificial intelligence. Automatic speech recognition has become an indispensable part of many everyday applications. Only the precise translation of speech to text with seemingly unlimited vocabulary enabled the widespread application of speech assistants in smartphones, smart speakers, and other devices. For many applications, speech recognition systems achieve a recognition accuracy close to humans. However, automatic speech recognition is not a solved problem. There is still no system robustly providing high transcription accuracy in all possible domains, applications, and speech situations. This statement is particularly true for languages other than English, where an enormous amount of transcripted speech is used for training.

One application that still poses enormous challenges for automatic transcription systems is *oral history*. In the humanities, oral history refers to conducting and analyzing interviews with contemporary witnesses to historical events. Oral history archives are often large audiovisual data repositories composed of numerous interviews, often a few hours in length per interview. Despite the enormous progress in speech recognition, for many archives, the transcription accuracy of interviews is still in a range that, at best, allows for a keyword-based search in transcripts. Therefore, the transcription is still performed entirely by humans in many archives. Oral history interviews pose a great challenge for automatic transcription systems due to their heterogeneity in terms of language, recording quality, speech type, dialects, speaker characteristics, and more. A lack of suitable, representative training data further limits the development of robust systems for interview transcription.

Many oral history interviews were recorded decades ago, and many interviewees have passed away in the meantime, leaving only their interview recording as a legacy for posterity and historical research. Therefore, the automatic transcription of these interviews is of great interest, not only because of the technical challenges. Automatic transcription can facilitate the work of historians, making interviews accessible and searchable for posterity and the research community in the form of time-aligned transcripts and subtitles. The automatic transcription of archives could even open up new research opportunities.

The presented research in this thesis aims to develop and improve robust speech recognition for German oral history interviews of a large archive. We explore

and study several approaches for the adaptation to improve systems' transcription accuracy and robustness and overcome the lack of training data. In the following, we first give a brief historical outline of research on automatic speech recognition in general to position robust speech recognition for oral history interviews in this vast field and put the challenge into perspective to past and current tasks.

1.1 Historical Outline of Automatic Speech Recognition

The first basic automatic speech recognition systems date back to the mid-twenties. These systems could recognize a very limited vocabulary such as digits or syllables, usually for a single speaker. One example of such a system that could recognize spoken digits is the *Audrey* system by Davis et al. [1952] from the Bell Laboratories.

The widespread application of *hidden Markov models* (HMM) in the 1980s to model speech with a statistical approach, e.g., by Levinson et al. [1983], can be regarded as the beginning of speech recognition as we understand it today. These approaches shaped the research and development of speech recognition systems for many decades. With the continuously improving recognition performance of proposed systems, increasingly challenging tasks became the focus of research and development.

Figure 1.1 by Huang et al. [2014, p. 96] illustrates the historical progress from 1988 to 2006 for various well-known English continuous speech recognition tasks. In the early 1990s, the primary challenge was speech recognition of read-aloud texts (*read speech*) with a very limited vocabulary of 1000 different words. In 1992, the Wall-Street-Journal-based continuous read speech recognition corpus [Paul and Baker, 1992] was published and became the focus of ASR research. The corpus proposed two tasks, one with a vocabulary of 5000 different words and one with 20,000.

Robust speech recognition, i.e., the automatic recognition of speech distorted by different acoustic effects, was considered in the mid-1990s and then gained attention again with the Aurora4 task [Parihar and Picone, 2002] (not shown in the graph). In this task, various noise classes from real-life recordings were added to the Wall Street Journal corpus to simulate realistic speech recognition challenges. According to Benesty et al. [2008, p. 654] and Li et al. [2014], the Aurora4 task can be considered a standard noise-robustness large vocabulary continuous speech recognition (LVCSR) task.

The Wall Street Journal and Aurora4 tasks were proposed as large vocabulary tasks. However, from today's perspective, where ASR systems recognize several hundred thousand to several million different words, a few-thousand-word vocab-

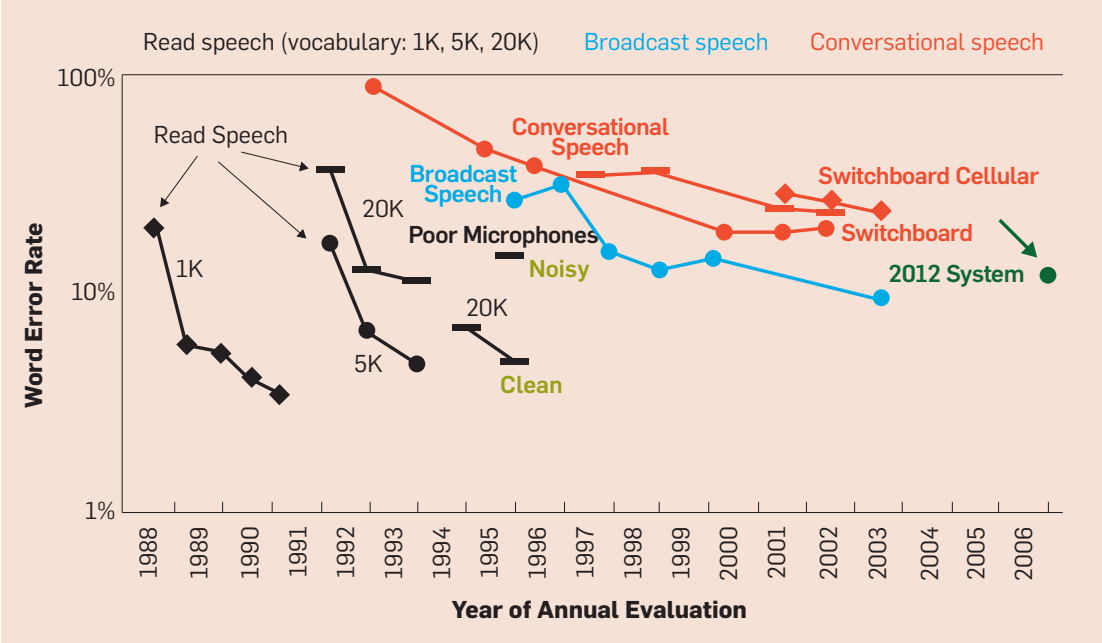


Figure 1.1: Historical progress of word error rates on more and more challenging speech recognition tasks [Huang et al., 2014, p. 96].

ulary task is considered relatively simple. As shown in one of our primary works [Hirsch and Gref, 2017], very low word error rates can be achieved on the Aurora4 task with relatively simple neural network architectures, both by speech enhancement and multi-condition training.

In the late nineties, the focus shifted from read speech to more challenging broadcast and conversational speech tasks. In particular, the Switchboard *English Conversational Telephone Speech Recognition* task (part of the NIST 2000 Hub-5 benchmark) was regularly used until the late 2010s by major research laboratories (such as IBM T. J. Watson Research Center [Saon et al., 2015], [Saon et al., 2016], [Saon et al., 2017] and Microsoft Research [Xiong et al., 2017b], [Xiong et al., 2018]) to competitively evaluate the performance of English ASR systems. The corpus consists of recordings of telephone conversations between strangers about random topics.

Hidden Markov models were extensively used to build ASR systems until the 2010s. Then, empowered by the increasing popularity of the deep learning paradigm, *deep neural networks* (DNN) were applied more and more frequently in automatic speech recognition for acoustic modeling. Often, DNNs were used to extend the conventional HMM approach in a *hybrid DNN-HMM* system. Hinton et al. [2012] give an overview of the progress and developments of such hybrid DNN-HMMs at that time.

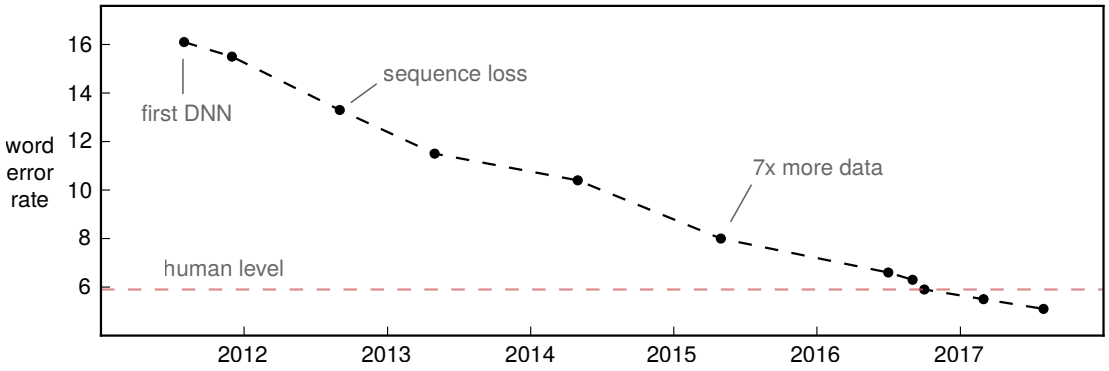


Figure 1.2: Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark.[1]

The application of neural networks for acoustic modeling in automatic speech recognition is not entirely new. For example, in the survey given by Trentin and Gori [2001], in the time before the deep learning paradigm, many systems and approaches have been described that combine hidden Markov models and artificial neural networks for acoustic modeling. Some of these systems have already used *recurrent neural networks* and *time delay neural networks*—architectures *rediscovered* for acoustic modeling by recent research and applied in ASR systems enabling cutting edge results. However, it took time until the deep learning paradigm for neural-network-based systems regularly outperformed traditional GMM-HMM-based systems. It was observed that neural networks generalize better than GMM-HMM-based models being trained with the same large amounts of training data. Therefore, it is not surprising that nowadays, the amount of training data plays an enormous role in developing modern speech recognition systems, and models are trained on a vast amount of annotated speech. For instance, Soltau et al. [2017] trained a model on 125,000 hours of annotated speech, i.e., more than 14 years of non-stop speech, for Google's speech recognition.

Through the advances of DNNs, automatic speech recognition research gained new momentum that continues until today. The impact of neural networks on speech recognition becomes clear considering the results of speech recognition systems on the aforementioned switchboard task in recent years. As an example, Figure 1.2 shows the results of some ASR systems between 2012 and 2017, which were state-of-the-art at the respective time.

In 2017, Saon et al. [2017] from *IBM Watson* achieved 5.5 % word error rates on the Switchboard task. This word error rate is considered by some researchers, e.g., by Xiong et al. [2017a], to be at the same level as professional human transcribers. Parallel to these advances, new challenges emerged, particularly in recognition of

[1] Source: https://awni.github.io/speech-recognition

spontaneous, informal speech, speech recorded under difficult acoustic conditions, dialects, and low-resourced languages, to name but a few. This is partly due to the increasing popularity of speech assistants at that time, such as Amazon's Alexa, Google Home, Microsoft's Cortana, and Apple's Siri, and the associated growing expectations of customers for the systems to work robustly in all situations and for every language.

Therefore, recent ASR tasks explore these new scenarios, such as the *dinner party scenario* of the 5th CHiME Speech Separation and Recognition Challenge (CHiME-5) [Barker et al., 2018]. This challenge addresses the combined problem of distant multi-microphones, spontaneous conversations in the presence of background noise, and simultaneously talking speakers. Even cutting-edge systems in 2018 achieved a word error rate higher than 45 % [Du et al., 2018], [Kanda et al., 2018] on this task.

In particular, these newer challenges are, at least in some aspects, similar to oral history interviews as they study poor recording conditions and spontaneous speech. However, the challenges are quite limited, and the conditions are much more predictable than for oral history. Hardly any common ASR task models challenges in a way comparable to oral history interviews. As we discuss in Section 3.3, for twenty years, researchers studied automatic transcription of oral history interviews for various languages. However, many works are usually characterized by a much higher error rate than common ASR challenges, leaving automatic transcription of oral history an unsolved issue.

1.2 List of Key Contributions

List of key contributions in this thesis:

- A human word error rate on German oral history interviews in clean acoustic conditions is experimentally estimated in Chapter 3.
- Noise and reverberation data augmentation is proposed and studied to improve the real-world performance of LF-MMI acoustic models for oral history interviews and other domains with unseen conditions by reducing the domain mismatch and improving acoustic robustness in Chapter 4.
- Two-staged LF-MMI acoustic model domain adaptation is proposed and investigated, combining data augmentation for acoustic robustness with acoustic model fine-tuning in Chapter 5. The approach is studied using a leave-one-speaker-out cross-validation and semi-automatically created adaptation data using automatic transcript alignment.

- Multi-staged cross-lingual adaptation is proposed and investigated that reduces domain overfitting and increases the robustness of the domain-adapted LF-MMI acoustic model with a cross-lingual pre-training stage in Chapter 6.

Further contributions are summarized at the end of each chapter.

1.3 List of Publications

Parts of this thesis have been published in peer-reviewed journals and conference proceedings. The most relevant publications covering the main chapters of this thesis are presented below in chronological order:

> Michael Gref, Joachim Köhler, and Almut Leh. Improved transcription and indexing of oral history interviews for digital humanities research. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 3124–3131. European Language Resources Association (ELRA), 2018a. URL https://aclanthology.org/L18-1493

> Michael Gref, Christoph Schmidt, and Joachim Köhler. Improving robust speech recognition for German oral history interviews using multi-condition training. In *13th ITG Conference on Speech Communication*, pages 256–260. VDE / IEEE, 2018b. URL https://ieeexplore.ieee.org/document/8578034

> Michael Gref, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Two-staged acoustic modeling adaption for robust speech recognition by the example of German oral history interviews. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 796–801, 2019. doi:10.1109/ICME.2019.00142

> Michael Gref, Oliver Walter, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Multi-staged cross-lingual acoustic model adaption for robust speech recognition in real-world applications - A case study on German oral history interviews. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6354–6362. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.780

> Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*, 2022b. URL https://arxiv.org/abs/2201.06841

The following publications (in chronological order) related to the thesis' research topic were written with the thesis author's contribution as a co-author. These are cited in the present research work as external literature and do not cover significant parts of the chapters:

Joachim Köhler, Michael Gref, and Almut Leh. KA3. Weiterentwicklung von Sprachtechnologien im Kontext der Oral History. *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, Schwerpunktheft: Digital Humanities und biographische Forschung*, 30(1-2/2017):44–59, 2019. doi:10.3224/bios.v30i1-2.05

Almut Leh, Joachim Köhler, Michael Gref, and Nikolaus Himmelmann. Speech analytics in research based on qualitative interviews. experiences from KA3. *VIEW Journal of European Television History and Culture*, 7(14):138–149, 2018. doi:10.18146/2213-0969.2018.jethc158

Almut Leh, Michael Gref, and Joachim Köhler. Audio mining. advanced speech analytics for oral history. *Words and Silences/Palabras y Silencios*, (2018-2019):1–9, 2019. URL https://www.ioha.org/wp-content/uploads/2019/10/Leh_IOHA_2018_Audiomining_English.pdf

Jan Gorisch, Michael Gref, and Thomas Schmidt. Using automatic speech recognition in spoken corpus curation. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6423–6428. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.790

Yao Wang, Michael Gref, Oliver Walter, and Christoph Schmidt. Bilingual i-vector extractor for DNN hybrid acoustic model training in German speech recognition systems. In *14th ITG Conference on Speech Communication*, pages 29–33. VDE / IEEE, 2021. URL https://ieeexplore.ieee.org/document/9657501

Michael Gref, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke, and Joachim Köhler. A study on the ambiguity in human annotation of german oral history interviews for perceived emotion recognition and sentiment analysis. In *13th International Conference on Language Resources and Evaluation (LREC)*, pages 2022–2031. European Language Resources Association (ELRA), 2022a. URL https://aclanthology.org/2022.lrec-1.217

The following publications (in chronological order) related to speech processing and automatic speech recognition were written with the thesis author's contribution as a co-author during the presented research. These are cited in the thesis as external literature and do not cover significant parts of the chapters:

> Hans-Günter Hirsch and Michael Gref. On the influence of modifying magnitude and phase spectrum to enhance noisy speech signals. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1978–1982, 2017. doi:10.21437/Interspeech.2017-1173

> Hans-Günter Hirsch and Michael Gref. Keyword detection for the activation of speech assistants. In *13th ITG Conference on Speech Communication*, pages 186–190. VDE / IEEE, 2018. URL https://ieeexplore.ieee.org/document/8578020

> Hans-Günter Hirsch, Alexander Micheel, and Michael Gref. Keyword detection for the activation of speech dialogue systems. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung (ESSV)*, pages 2–9. TUDpress, Dresden, 2020. ISBN 978-3-959081-93-1. URL https://www.essv.de/paper.php?id=431

> Julia Pritzen, Michael Gref, Christoph Schmidt, and Dietlind Zühlke. A comparative pronunciation mapping approach using G2P conversion for anglicisms in German speech recognition. In *14th ITG Conference on Speech Communication*, pages 24–28. VDE / IEEE, 2021. URL https://ieeexplore.ieee.org/document/9657500

> Julia Pritzen, Michael Gref, Dietlind Zühlke, and Christoph Andreas Schmidt. Multitask learning for grapheme-to-phoneme conversion of anglicisms in German speech recognition. In *13th International Conference on Language Resources and Evaluation (LREC)*, pages 3242–3249. European Language Resources Association (ELRA), 2022. URL https://aclanthology.org/2022.lrec-1.346

1.4 Public Model Access

We provide public access for academic research to the best performing model from Chapter 4 with a free limited monthly contingent as part of the *BAS Speech Science Web Services* [Kisler et al., 2016].[2] The released model is adapted to challenging acoustic conditions in oral history interviews and other real-world speech recordings and uses a general-purpose broadcast language model.

[2] https://clarin.phonetik.uni-muenchen.de/BASWebServices

1.5 Thesis Outline

This thesis is structured as follows:

Chapter 2 presents backgrounds and related works in the field of automatic speech recognition. In particular, it focuses on methods and state-of-the-art approaches relevant to the presented research work.

Chapter 3 discusses the automatic transcription of oral history interviews, reviewing related work on this topic, and analyzes the challenges of oral history interviews for speech recognition. In addition, human transcription is analyzed, and preliminary experiments are conducted to identify the components of speech recognition systems to be improved.

Chapter 4 investigates approaches to improve the robustness of acoustic models against challenging recording conditions commonly encountered in oral history interviews.

Chapter 5 explores approaches for adapting acoustic models to the oral history domain.

Chapter 6 studies multi-stage cross-lingual adaptation of acoustic models to exploit the vast amount of English ASR corpora to improve speech recognition for German oral history.

Chapter 7 concludes the thesis and summarizes the results and approaches.

2 Automatic Speech Recognition

Automatic speech recognition is a research field that aims at processing human speech signals in order to automatically obtain spoken words in the form of a written text. Most modern speech recognition systems are statistical approaches that estimate the probability of word sequences for *observed features* extracted from a given discrete-time speech signal. Let $\hat{\boldsymbol{w}} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N) \in \mathcal{W}$ be a word sequence from a set of all possible (finite) word sequences $\mathcal{W}$ and let $\boldsymbol{X}$ be the observed speech features. Usually, $\boldsymbol{X}$ is an arbitrarily long sequence of feature vectors, i.e., $(\boldsymbol{x}_n)_{n=1}^T$, extracted from the raw speech signal. The probability estimated by a speech recognition system for word sequence $\boldsymbol{w}$ subject to a given feature $\boldsymbol{X}$ is

$$P_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{X})$$

where $\boldsymbol{\theta}$ is a tuple of all parameters of the speech recognition system.

During training, the aim is to obtain model parameters such that the estimated probability matches the training data best. For a trained model, the model parameters are usually kept fixed. Therefore, we usually omit $\boldsymbol{\theta}$ for a clearer, simpler notation when we consider trained models.

The task of obtaining the most probable word sequence $\hat{\boldsymbol{w}} \in \mathcal{W}$ given a feature $\boldsymbol{X}$ using a trained speech recognition system is called *decoding*. Mathematically, cf. Yu and Deng [2015, pp. 101–102], this can be written as

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \mathcal{W}} P(\boldsymbol{w}|\boldsymbol{X}), \tag{2.1}$$

i.e., finding the word sequence $\boldsymbol{w}$ of all possible word sequences $\mathcal{W}$ for which the probability estimated by the model is highest. Many decades of research have shown that it is hard to design and train a system that sufficiently estimates $P(\boldsymbol{w}|\boldsymbol{X})$ straightforwardly. Instead, to estimate this probability, for many decades, approaches became state-of-the-art that utilize the well-known *Bayes rule* and are based on certain assumptions about the probabilities of speech.

This chapter gives an overview of the different theoretical fundamentals of common approaches in the speech recognition field. Section 2.1 discusses automatic speech recognition based on conventional hidden Markov models that dominated the field for many decades and that are still a component of recent speech recognition systems. The *hybrid* combination of hidden Markov models with deep neural networks is discussed in Section 2.2. In Section 2.3, we discuss the emerging re-

search field of *end-to-end speech recognition* that performs the entire recognition using a single sequence-to-sequence deep neural network. Sequence discriminative training that incorporates ideas of end-to-end speech recognition for hybrid models is discussed in Section 2.4.

2.1 Automatic Speech Recognition Using Hidden Markov Models

In the mid-1970s, researchers began to process continuous speech statistically using *hidden Markov models* (HMMs) and applied this approach to automatic speech recognition (e.g., Baker [1975]). In the early 1980s, hidden Markov models have established state-of-the-art and were applied for most speech recognition systems. Hidden Markov models were used for the recognition of isolated words, e.g., by Levinson et al. [1983], connected words, e.g., by Rabiner and Levinson [1985], and continuous speech recognition, e.g., by Bahl et al. [1983]. This can be regarded as the beginning of our current understanding of continuous automatic speech recognition. For many decades, hidden Markov models enabled the development of cutting-edge speech recognition systems, like Soltau et al. [2005] as one example for such works before deep neural networks emerged in automatic speech recognition.

2.1.1 Theory of Hidden Markov Models

The basic mathematical theory behind hidden Markov models was published in a series of papers by Baum et al. in the late 1960s and early 1970s: [Baum and Petrie, 1966], [Baum and Eagon, 1967], [Baum and Sell, 1968], [Baum et al., 1970] and [Baum, 1972]—cf. Rabiner [1989]. At its core, a hidden Markov model describes two associated stochastic processes:

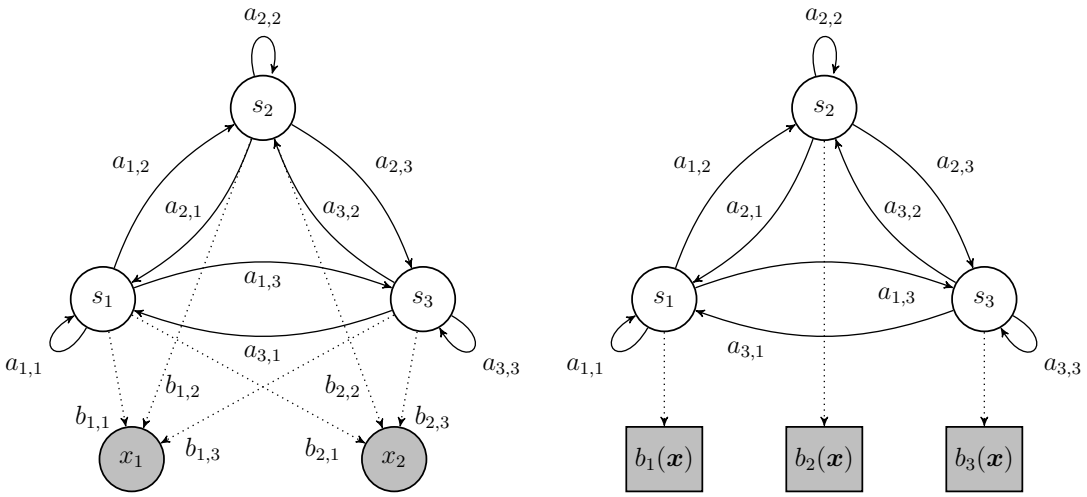
1. a Markov chain where the current state during a sequence is not directly visible for an observer (i.e., *hidden*) and

2. a second stochastic process that is observable and whose current outputs depend on the current state of the (hidden) Markov chain.

In other words, hidden Markov models extend conventional discrete-time Markov processes with a second, stochastic process that generates an *observable* sequence

$$\boldsymbol{o} := (o_t)_{t \in \{1,2,\ldots,T\}}.$$

from a set of observable events X, i.e., $o_t \in X$, as a probabilistic function of the state. This sequence is called *observable* in contrast to the underlying (hidden)



(a) Discrete observation distribution. Two observations are represented as gray circles. Each state has exactly one probability for each of the observations.

(b) Continuous observation distribution. Three probability density functions of observations, each associated with one state, are represented as rectangles.

Figure 2.1: Schematic representations of general hidden Markov models with three hidden states and the different types of observation distributions. Boxes and circles with gray backgrounds represent the observation distributions. Dotted lines represent the association of states to observation probabilities.

Markov sequence $\boldsymbol{q}$ that is not unambiguously determinable for any given sequence of observations. Generally, two different cases of observations are considered: *discrete (and finite) observation distributions* and *continuous observation distribution*. In case of discrete observation distribution, the amount of distinguishable possible observations $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\} \subset \mathbb{R}^D$ is finite, i.e., the probability of observing $\boldsymbol{x}_m$ at state s_k is defined as

$$b_{m,k} := P(o_t = \boldsymbol{x}_m | q_t = s_k).$$

In the case of continuous observation distributions, the probability of observing event $\{\boldsymbol{x}\} \subseteq \mathbb{R}^D$ at state s_k is usually modeled using probability density functions $b_k \colon \mathbb{R}^D \to \mathbb{R}, \boldsymbol{x} \mapsto b_k(\boldsymbol{x})$. Each function b_k is associated with one state s_k. A schematic representation of both types of hidden Markov models is given in Figure 2.1

Hidden Markov models with continuous observation distributions are the type of models usually used for automatic speech recognition. Hence, we focus on the theory of such models in the following. A general representation of a probability density function, cf. Rabiner [1989] and similarly Yu and Deng [2015, p. 28], and

Vasquez et al. [2012, p. 27], can be approximated by a finite mixture of form

$$b_k(\boldsymbol{x}) = \sum_{m=1}^{M} c_{k,m} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m})$$

with Gaussian mean-vector $\boldsymbol{\mu}_{k,m} \in \mathbb{R}^D$, covariance-matrix $\boldsymbol{\Sigma}_{k,m} \in \mathbb{R}^{D \times D}$ and mixture gains $c_{k,m} \in \mathbb{R}$ that satisfy

$$\forall k \in \{1, ..., N\} : \sum_{m=1}^{M} c_{k,m} = 1$$

so that

$$\forall k \in \{1, ..., N\} : \int_{\mathbb{R}^D} b_k(\boldsymbol{x}) d\boldsymbol{x} = 1$$

is fulfilled and b_k is a well-defined probability density function.

Since the mixture is based on Gaussian distributions $\mathcal{N}$, the approach is called *Gaussian mixture model* (GMM). A hidden Markov model using Gaussian mixture models for observation probabilities is called *Gaussian mixture model - hidden Markov model* (GMM-HMM).

In automatic speech recognition, hidden Markov models are applied to model humans' speech production process statistically. This can be *speaker-dependent*, i.e., a finite set of models that statistically model the speech of one specific speaker, or *speaker-independent*, i.e., a finite set of models that model the speech production of arbitrary speakers. Depending on the desired application, cf. Vasquez et al. [2012, p. 23], the hidden Markov chain of a single hidden Markov model can represent either an entire word, a phone, or *context-dependent* phones (e.g., *biphones* or *triphones*). The latter take into account that the sound of phones can change depending on the succeeding or preceding phone and therefore use a different hidden Markov model for each possible combination of context-dependent phones.

Triphones are a suitable and popular approach to reliably model the context-dependency of phones in large vocabulary continuous speech recognition systems. However, due to the enormous amount of possible triphone combinations for a given phone set, many triphones often do not appear in the given training data or appear too seldom to be learned reliably by a hidden Markov model, cf. Beulen et al. [1997]. To reduce complexity and the total number of context-dependent hidden Markov models, in modern systems, often *state tying* [Beulen et al., 1997] is performed. This approach clusters similar observation probability distributions of states across different Markov models, i.e., distributions that model similar acoustic events. Such states that share a clustered distribution are called *tied*

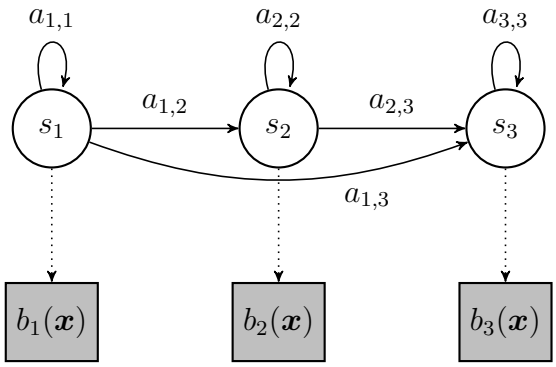


Figure 2.2: Schematic representation of a left-right hidden Markov model with three hidden states and continuous observation distributions (3-state Bakis model). Each rectangle represents a probability density function of observations associated with one state.

states. Single states in such systems can be considered *subphonetic* units, also called *senones*, cf. Hwang and Huang [1992].

Considering the temporal causality of speech production, in practice, usually, *left-to-right models* or *Bakis models* [Bakis, 1976] are used that do not allow state-transitions backward in time, e.g., as shown in Figure 2.2.

2.1.2 Automatic Speech Recognition Based on Hidden Markov Models

Hidden Markov models are considered generative models. Their popularity in automatic speech recognition stems from their ability to model acoustic features of speech [Yu and Deng, 2015, p. 42]. Thus, as generative models, hidden Markov models cannot estimate $P(\boldsymbol{w}|\boldsymbol{X})$, i.e., the probability of word sequence $\boldsymbol{w}$ subject to a given feature sequence $\boldsymbol{X}$, as specified in the speech recognition decoding problem formulation (Equation 2.1), in a direct manner. However, using Bayes' rule, the equation is rewritten as follows

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}\in\mathcal{W}} P(\boldsymbol{w}|\boldsymbol{X}) = \arg\max_{\boldsymbol{w}\in\mathcal{W}} \left(\frac{P(\boldsymbol{X}|\boldsymbol{w}) \cdot P(\boldsymbol{w})}{P(\boldsymbol{X})} \right)$$

in case $P(\boldsymbol{X}) \neq 0$. Since $P(\boldsymbol{X}) > 0$ is independent from $\boldsymbol{w}$, the recognized word sequence $\hat{\boldsymbol{w}}$ is obtained from

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}\in\mathcal{W}} (P(\boldsymbol{X}|\boldsymbol{w}) \cdot P(\boldsymbol{w})), \tag{2.2}$$

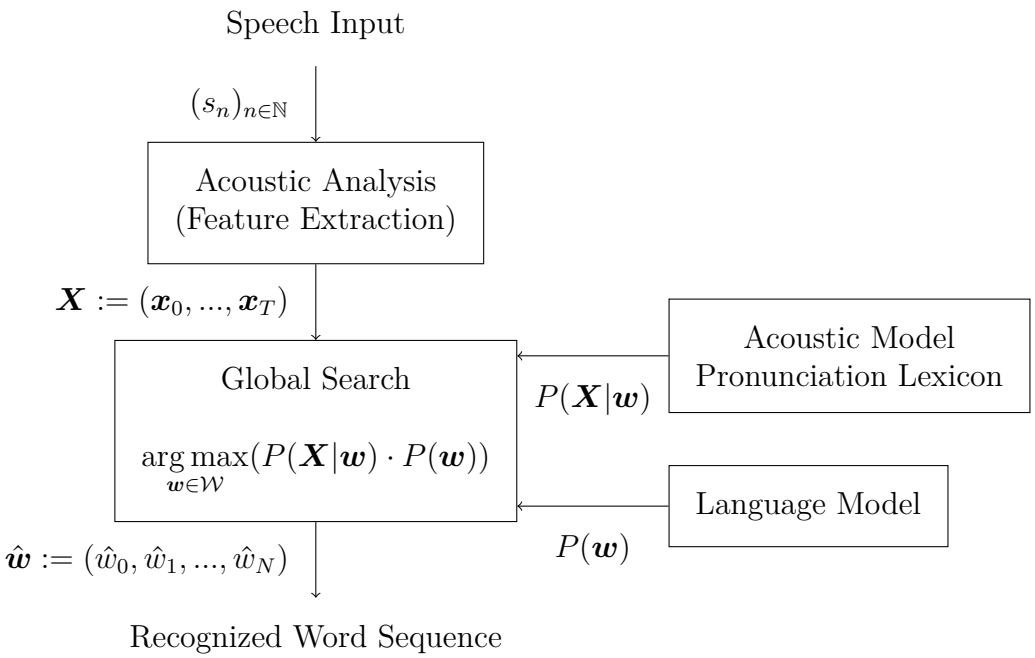


Figure 2.3: Bayes' decision rule for stochastic automatic speech recognition, cf. Ney and Ortmanns [1999].

cf. Jelinek [1976]; Bahl et al. [1983]. In this context, the model used to estimate $P(\boldsymbol{w})$—the probability of word sequence $\boldsymbol{w}$—is called *language model* (LM). Probability $P(\boldsymbol{X}|\boldsymbol{w})$, i.e., observing feature sequence $\boldsymbol{X}$ given word sequence $\boldsymbol{w}$, is estimated by a component called *acoustic model* (AM) along with—in case of phone-based speech recognition—a *phonetic pronunciation lexicon*. A popular schematic representation of the speech recognition decoding process based on Equation 2.2 (*Bayes' decision rule*, cf. Ney and Ortmanns [1999]), is shown in Figure 2.3.

In large vocabulary continuous speech recognition, usually one 3-state Bakis HMM models one subword unit—a single phone or context-dependent phone via tied states. Multiple *subword hidden Markov models* are concatenated to form *word hidden Markov models*. These can model arbitrary, continuous word sequences of unknown length $N \in \mathbb{N}$. An example for a word model, based on concatenated phone hidden Markov models, is presented in Figure 2.4. These word models are based on the phonetic transcriptions of the respective word defined in the pronunciation lexicon.

Word hidden Markov models are then further concatenated, enabling transitions from the end state of every word to the initial state of every word, cf. Ney and Ortmanns [1999]—as illustrated in Figure 2.5 for a simple example. Optional pause

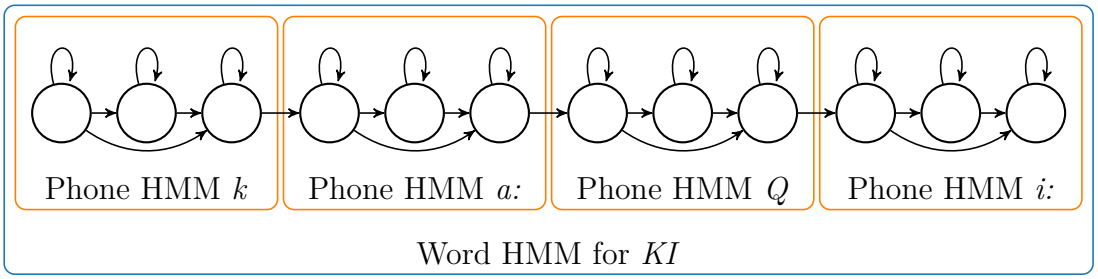


Figure 2.4: Illustration of concatenated phone hidden Markov models that model the pronunciation of the German word *KI* (abbreviation for *künstliche Intelligenz*, German for *artificial intelligence*). The BasSAMPA phone set is used for phonetic transcription. Visualizations of hidden Markov model observation probability distributions have been omitted in this figure for better clarity.

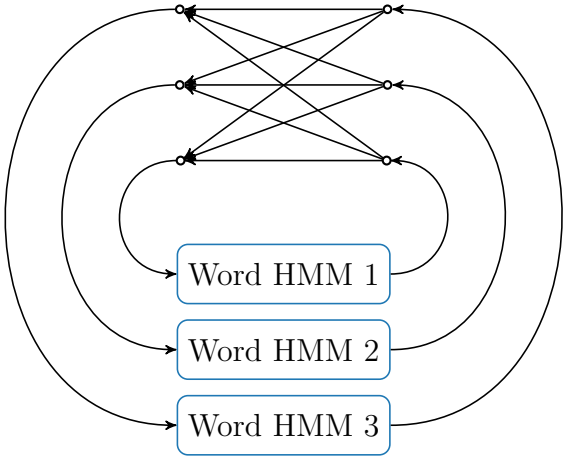


Figure 2.5: Simplified illustration of an HMM acoustic model with a three-word vocabulary, cf. Ney and Ortmanns [1999]. Each blue block represents an entire (left to right) word hidden Markov model as in Figure 2.4.

models between these word models can model naturally occurring speech pauses between words. Concatenating hidden Markov models in such a way results in a huge word sequence hidden Markov model. It approximately models the entire speech production process for arbitrary word sequences via observation probabilities associated with the sequences of subword hidden Markov states $\boldsymbol{s}$ in a huge finite network—i.e., probability $P(\boldsymbol{X}|\boldsymbol{w})$. The set of all GMM-HMMs applied in such a network is called *(GMM-HMM) acoustic model*.

From the stochastic point of view, the relation between the acoustic model and the pronunciation lexicon in the decoding process can be described as follows. In order to recognize the most probable word sequence according to Bayes' decision rule (Equation 2.2), probability $P(\boldsymbol{X}|\boldsymbol{w})$ is factorized further to products of probabilities that the acoustic model estimates. Applying a hidden Markov acoustic

model, the probability $P(\boldsymbol{X}|\boldsymbol{w})$ is equal to the sum of all probabilities generating observation sequence $\boldsymbol{X} := (\boldsymbol{x}_0, ..., \boldsymbol{x}_T)$ of length $T \in \mathbb{N}$ for all possible subword state sequences $\boldsymbol{s} := (s_0, ..., s_T) \in \mathcal{S}$ of length T, i.e.,

$$P(\boldsymbol{X}|\boldsymbol{w}) = \sum_{\boldsymbol{s}\in\mathcal{S}} P(\boldsymbol{O} = \boldsymbol{X}, \boldsymbol{q} = \boldsymbol{s}|\boldsymbol{w}),$$

or in short notation

$$P(\boldsymbol{X}|\boldsymbol{w}) = \sum_{\boldsymbol{s}\in\mathcal{S}} P(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{w}).$$

This equation can be further factorized

$$P(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{w}) = P(\boldsymbol{X}|\boldsymbol{w}, \boldsymbol{s}) \cdot P(\boldsymbol{s}|\boldsymbol{w})$$

using Kolmogorov's definition of conditional probability. With a conditional independence assumption for the acoustic model, cf. Watanabe et al. [2017]; Wang et al. [2019], we approximate $P(\boldsymbol{X}|\boldsymbol{w}, \boldsymbol{s}) \approx P(\boldsymbol{X}|\boldsymbol{s})$. Thus, the (general) Bayes' decision rule is further factorized as follows

$$\hat{\boldsymbol{w}} \approx \arg\max_{\boldsymbol{w}\in\mathcal{W}} \left(P(\boldsymbol{w}) \cdot \sum_{\boldsymbol{s}\in\mathcal{S}} P(\boldsymbol{X}|\boldsymbol{s}) \cdot P(\boldsymbol{s}|\boldsymbol{w}) \right) \tag{2.3}$$

cf. Watanabe et al. [2017]; Wang et al. [2019]. For this *factorized Bayes' decision rule*, $P(\boldsymbol{X}|\boldsymbol{s})$ is estimated using the observation probabilities of the acoustic model—e.g., the Gaussian mixture models in a GMM-HMM setup. Probability $P(\boldsymbol{w})$ is still estimated by the language model, as described above. Probability $P(\boldsymbol{s}|\boldsymbol{w})$ describes the probability of subword state sequence $\boldsymbol{s}$ given word sequence $\boldsymbol{w}$. It is modeled by the *pronunciation model* (or *lexicon model*), cf. Watanabe et al. [2017]; Wang et al. [2019], that includes the HMM state transition and initial state-occupation probabilities of the acoustic model, the phonetic pronunciation lexicon entries as well as mappings between the state sequence units (such as tied states) and the phonetic units used in the dictionary. Optional silence and pronunciation probabilities [Chen et al., 2015] might also be used in this model to improve the estimation of $P(\boldsymbol{s}|\boldsymbol{w})$.

Some works, such as the aforementioned works by Watanabe et al. [2017]; Wang et al. [2019] refer to the components that estimate $P(\boldsymbol{X}|\boldsymbol{s})$ as the acoustic model. However, this definition may be misleading, as it contradicts the above definition respectively only covers a subset of the above characterization of a GMM-HMM acoustic model. The definition in the aforementioned works only considers the observation probabilities a part of the acoustic model but not the transition and initial state occupation probabilities of the hidden Markov model used in the pro-

nunciation model. Therefore, we refrain from this definition and adhere to the term *(acoustic) observation probabilities* for the components estimating $P(\boldsymbol{X}|\boldsymbol{s})$.

In applications, calculating the entire sum over all possible state sequences of all probabilities in Equation 2.3 is not feasible for large-scale vocabulary continuous speech recognition systems. Instead, often *Viterbi approximation*, cf. Ney and Ortmanns [1999], is applied in Bayes' decision rule. Viterbi approximation substitutes the sum with the max operator over all possible state sequences probabilities in the decision rule, i.e.,

$$\hat{\boldsymbol{w}} \approx \arg\max_{\boldsymbol{w}\in\mathcal{W}} \left(P(\boldsymbol{w}) \cdot \max_{\boldsymbol{s}\in\mathcal{S}} (P(\boldsymbol{X}|\boldsymbol{s}) \cdot P(\boldsymbol{s}|\boldsymbol{w})) \right). \tag{2.4}$$

It is noteworthy that this does not mean that the Viterbi approximation and full-sum probabilities are similar but that the arg max search often leads to similar recognition results. The maximization, however, facilitates efficient and fast implementations for hidden Markov models.

The factorization of Bayes' decision rule for HMM-based speech recognition systems allows splitting the system into three different models: the acoustic observation probabilities, the language model, and the pronunciation model.[1] The decomposition of phone-based speech recognition systems into these three individual components has a substantial advantage over alternative approaches. The language and acoustic model can be trained independently using entirely different types of data. Annotated speech is required to train the acoustic model—which is often difficult to obtain. To train a language model, however, only text data is needed. Such data is usually available in large amounts for different domains—such as in digitized books, news, websites, and many more. Furthermore, pronunciations of words can be modeled and adjusted independently from these two models.

Since the acoustic model in such a setup is trained to model phonetic or subphonetic units only—independently of the respective words in the training data—an already trained acoustic model is often also used to recognize many new words that did not appear in the training data for the acoustic model. The respective word must only appear in the training data for the language model.

In practice, the decoupling of the acoustic and language model has the great advantage of adapting systems with little effort to work well in different language domains. By substituting the language model, speech recognition systems can be set up for different domains using the same acoustic model. Such domains can be, for instance, the recognition of speech in news shows, the recognition of command

[1] An extension of Figure 2.3 which incorporates these changes is shown in Figure B.1 in the appendix. A simplified, informal schematic structure of the components and their interconnection in a large-vocabulary automatic speech recognition system is presented in Figure B.2.

words in a home control system, or the recognition of telephone numbers in a call center.

2.1.3 Features for Automatic Speech Recognition

Mel-Frequency Cepstral Coefficients Features

As presented in the previous section, hidden Markov models have been successfully applied to model the human speech production process statistically in many speech recognition systems. Usually, spectral features, extracted from the produced speech signal, serve as observable features for hidden Markov models. Features based on the short-time Fourier-transformation ($\mathcal{STFT}$) have become widely established for speech processing. Such spectral features allow distinguishing different frequency components and acoustic characteristics of different speech sounds along the time axis.[2]

However, for hidden Markov models, the $\mathcal{STFT}$ of a signal is often considered too complex to be directly used as observable features. Not only is the result of the $\mathcal{STFT}$ complex-valued—due to Fourier transformation. The frequency dimension of the frames usually has to be chosen relatively high to uncover underlying frequency characteristics of the time signal. However, such a high-dimensional feature space might make it hard for a model to distinguish relevant and irrelevant information—such as each harmonic component's exact amplitude value and frequency bin. Furthermore, the frequency bins of a $\mathcal{DFT}$ frame are statistically correlated.

Diagonal covariance matrices are usually used to reduce the complexity of training GMM-HMMs, assuming that the elements of the feature vectors are uncorrelated.[3] Obviously, an $\mathcal{STFT}$ spectrogram violates this assumption. For these reasons, *Mel-frequency cepstral coefficients* (MFCCs) have become the quasi-standard for features in HMM-based speech recognition.

The *European Telecommunications Standards Institute* (*ETSI*) standardized feature extraction based on the *Mel-cepstrum* in ETSI ES 201 108 [2003] to ensure compatibility between terminals and a remote speech recognizer. Even though most systems use their own implementation of MFCC feature extraction, the fundamental principles are mainly the same. The algorithm defined in the aforementioned ETSI standard will be used to present the MFCC feature extraction

[2] See Figure B.3 in the appendix for an exemplary visualization and [Oppenheim et al., 1999, pp. 714–717] for further reading.

[3] It is noteworthy that generally, it is also possible to work with full covariance matrices, and many works studied approaches to reduce the computational load of working with full covariance matrices, e.g., by Bell and King [2009]. However, the potential performance gains do not seem to justify the drawbacks of such approaches. For most of today's applications, diagonal covariance matrices seem to have prevailed.

in the following. The basic steps required to extract MFCCs from a (sampled) discrete-time signal are schematically presented in Figure 2.6 in the form of a block diagram.

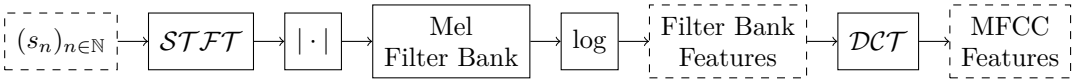


Figure 2.6: Workflow to obtain MFCC features.

After transforming the discrete-time signal using the short-time Fourier-transformation, the phase of the complex-valued matrix is removed by taking the absolute value element-wise. Subsequently, the feature dimension is reduced using a filter bank with the aim of optimizing them for speech recognition. For this purpose, human hearing is used as a model. Psychoacoustic studies show that humans have the highest frequency resolution at lower frequencies—between approximately 1 and 2 kHz, cf. Oxenham [2018]. At high frequencies, the resolution drastically decreases. The effect of non-linear frequency resolution of humans is also reflected in the pitch perceived by humans as a non-linear function of the frequency. The *Mel scale* [Stevens et al., 1937], cf. Figure B.4, is used to define the *Mel filter bank*, cf. Figure B.5, summarizing frequency intervals considered equally distant in human perception. The filter masks of this filter bank are applied to the frequency bins of each $|\mathcal{STFT}|$-frame to obtain N_F-dimensional vectors for each time step. As the next step, the natural logarithm is applied to the vector elements. This takes into account humans' non-linear perception of the loudness, which can be approximated as a non-linear, logarithmic relationship between the atmospheric pressure of sound waves and perceived loudness.

Finally, to obtain N_C-dimensional MFCCs feature vectors $\boldsymbol{c}$, the discrete cosine transformation ($\mathcal{DCT}$) is applied on the feature vectors of each frame to decorrelate the correlated features $\boldsymbol{z} := \left((z_n)_0^{N_F-1}\right)$ and further reduce feature dimension, i.e.,

$$\boldsymbol{c} := \mathcal{DCT}(\boldsymbol{z}) := \left(\sum_{n=0}^{N_F-1} z_n \cdot \cos\left(\frac{\pi \cdot k}{N_F}\left(n-\frac{1}{2}\right)\right)\right)_{k=0}^{N_C-1},$$

usually, with $N_C = 13$ for HMM-based systems.

It should be noted that the MFCC workflow defined in the ETSI standard includes some implementation details that were neglected in the above description to highlight the essential aspects of MFCC features that most common implementations share. For instance, the implementation also contains *pre-emphasis* or *offset compensation* designed as finite impulse response (FIR) filters.

Filter Bank Features

The values obtained right before the $\mathcal{DCT}$ step in the MFCC extraction workflow, as shown in Figure 2.6, are called *(Mel) filter bank features* or *fbank* features. The usage of filter bank features did not play a major role in GMM-HMM-based speech recognition. These features became more prominent with the emergence of more recent approaches that do not require uncorrelated features—such as neural networks that will be discussed in the following section. Many works, like Hinton et al. [2012], Deng et al. [2013], and Yoshioka et al. [2014], studied these features as inputs for neural-network-based speech recognition in different setups. Some results indicate that filter-bank features improve speech recognition performance compared to standard MFCC features—especially when the feature dimension is increased, e.g., to $N_F = 40$. However, other researchers, such as Povey et al. [2015], concluded that *high-resolution MFCCs* with $N_F = N_C = 40$ are equivalent to equal dimensional filter bank features because such MFCC features are a linear transform of the respective filter bank features.

Delta and Delta-Delta Features

As they are used for HMM-based speech recognition, Mel-frequency cepstral coefficient features analyze only very short segments of a speech signal. Therefore, the feature vectors do not contain any information beyond this short time horizon. As stated by Kumar et al. [2011], among others, *delta* and *delta-delta* features have been widely adopted in automatic speech recognition systems to add dynamic information of speech to the static cepstral features—at least before the emergence of deep learning in the speech recognition field. However, the fundamental concept of delta features is not limited to cepstral features. It can be reasonably applied to any type of features that are given as a temporal sequence of equidistant feature vectors—such as short-time spectrograms.

Delta features aim at approximating the first (element-wise) derivative of the underlying continuous function to the sampled sequence $(\boldsymbol{z}_n)_{n=0}^T$ of feature vectors. For $(\boldsymbol{z}_n)_{n=0}^T$, the delta feature at time step $k \in \{1, ..., T\}$ can be defined as

$$\Delta_k\Big((\boldsymbol{z}_n)_{n=0}^T\Big) := \boldsymbol{z}_k - \boldsymbol{z}_{k-1} \tag{2.5}$$

and initialization at $k = 0$. Respectively to delta features, delta-delta features (or *double-delta* features) aim at approximating the second derivative. Thus, delta-deltas are obtained by calculating the delta of delta-features, i.e, applying Equation 2.5 to the sequence of delta features:

$$\Delta\Delta_k\Big((\boldsymbol{z}_n)_{n=0}^T\Big) := \Delta_k\Big((\boldsymbol{z}_n)_{n=0}^T\Big) - \Delta_{k-1}\Big((\boldsymbol{z}_n)_{n=0}^T\Big).$$

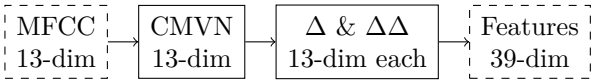


Figure 2.7: Conventional delta-delta MFCC features with CMVN for simple GMM-HMM ASR systems.

Figure 2.7 shows how MFCC delta-delta features are ultimately applied for speech recognition systems. Usually, first *cepstral mean and variance normalization* (*CMVN*) [Viikki and Laurila, 1998] (or *cepstral mean normalization*) is applied to the MFCC vectors. CMVN aims to provide zero-mean and unit-variance cepstral features—usually on per-speaker or per-utterance level—to improve the system's robustness to channel distortions.

Then the deltas and delta-deltas of the vectors are concatenated with the MFCC vector to obtain the final features. Nowadays, these types of features are usually only used for simple GMM-HMM systems, e.g., the first models in bootstrap training sequentially trained with Delta-Delta-MFCCs and used to generate alignments to train more advanced models. Much research has been carried out in the 1990s to study more sophisticated features and feature space transformations that overcome certain limitations of delta-delta features for large vocabulary GMM-HMM systems. These feature space transformations will be discussed in the following.

Linear Discriminant Analysis Feature Transformation

Haeb-Umbach and Ney [1992] studied *linear discriminant analysis* (*LDA*) for GMM-HMM-based large vocabulary speech recognition. LDA is a well-known approach in statistics based on Fisher's (linear) discriminant analysis [Mika et al., 1999] that aims at obtaining a linear transformation f of vectors $\boldsymbol{x}$ in the feature space $\mathbb{R}^N$ to feature vectors $\boldsymbol{y} := f(\boldsymbol{x}) \in \mathbb{R}^M$ with reduced dimension so that class separability is maximized [Haeb-Umbach and Ney, 1992]. Typically, for speech recognition, acoustic states—e.g., subphonetic units such as tied states—are used as classes to obtain the LDA transformation on *spliced* (concatenated $n \in \mathbb{N}$ consecutive) MFCC features.

Maximum Likelihood Linear Transformation

Another popular feature space transformation is *maximum likelihood linear transformation* (*MLLT*) [Gopinath, 1998]. In the literature, e.g., by Rath et al. [2013], MLLT is also known as *global semi-tied covariance* (*STC*) [Gales, 1999]. Unlike LDA, MLLT does not reduce the feature space dimension. Given the model parameters, the transformation is estimated to increase the likelihood of the training data and share Gaussian parameters across classes to improve discrimination ultimately. In practice, MLLT is often applied on top of LDA transformed features.

2.1.4 Speaker Adaptive Training of GMM-HMM Acoustic Models

For large vocabulary speech recognition systems, it is usually desired to train speaker-independent systems. Such systems are designed to be applied on arbitrary, unknown speakers that do not occur in the training data. Naturally, the great variability of different speakers poses a great challenge for the underlying acoustic models. *Speaker adaptive training* (SAT) aims to overcome this issue by reducing the speaker variability between training and test conditions, e.g., as proposed by Anastasakos et al. [1996]. The most common speaker adaptive training approaches for GMM-HMM-based acoustic models, cf. Yu and Deng [2015, pp. 163 f.], are *vocal tract length normalization* (VTLN) and *feature space maximum likelihood linear regression* (*fMLLR*) transform.

As the name implies, vocal tract length normalization aims to normalize variations in the speech signal of different speakers, such as varying formant frequencies in the spectrum caused by different vocal track lengths. Several implementations for VTLN have been studied and applied to speech recognition throughout the years, e.g., by Cohen et al. [1995], Eide and Gish [1996], Claes et al. [1998], Kim et al. [2004].

Feature space maximum likelihood linear regression is an affine transformation approach studied by Gales [1998] for speaker (and environmental) adaptation in GMM-HMM-based speech recognition systems. The approach aims at training an affine feature transform of form

$$f(\boldsymbol{x}) := \boldsymbol{A} \cdot \boldsymbol{x} + \boldsymbol{b}$$

with $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{b} \in \mathbb{R}^{D}$ that maximizes the likelihood of adaptation data given the model parameters.

Usually, fMMLR is applied on top of delta-delta-MFCC or LDA+MLLT transformed features, as described by Rath et al. [2013]. The LDA+MLLT+fMMLR approach, as depicted in Figure 2.8, aims to combine the advantages of the different presented transformations to obtain well separable, speaker normalized feature representations with reduced dimension. Other combinations of these and other transformations, such as VTLN, are also common. Such combined feature space transformations are more commonly used in more advanced large-vocabulary speech recognition systems.

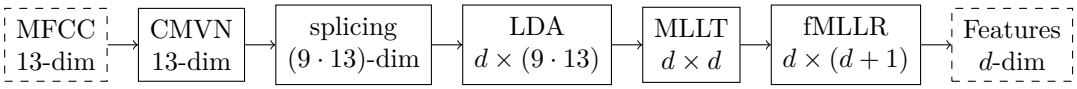


Figure 2.8: Advanced feature space transforms based on MFCCs for speaker adaptive training of GMM-HMM ASR systems. This figure is based on the baseline description in [Rath et al., 2013]. For LDA, nine consecutive MFCC feature vectors (i.e., with a temporal context of ± 4) are spliced (concatenated).

2.2 Hybrid Deep Neural Network - Hidden Markov Models

2.2.1 Neural Networks in HMM-based Speech Recognition

In the early 2010s, the deep learning paradigm found its way into speech recognition. Empowered by the performance of deep neural networks and the new opportunities that such models offer, automatic speech recognition regained enormous research interest by many academic research groups and companies in the last decade. Deep neural networks have been applied more and more often in speech recognition systems.

Until a few years ago, deep neural networks were usually applied to extend the acoustic modeling in conventional GMM-HMM-based speech recognition. Such systems are called *hybrid deep neural network - hidden Markov models* (*hybrid DNN-HMMs* or simply *DNN-HMMs*) [Yu and Deng, 2015, p. 99]. The deep neural network in such a hybrid system, as shown in Figure 2.9, is used to obtain a better estimation of the HMM acoustic model observation probabilities $P(\boldsymbol{X}|\boldsymbol{s})$ used for the (factorized) Bayes' decision rule (Equation 2.4). However, the neural network does not directly estimate $P(\boldsymbol{X}|\boldsymbol{s})$, since (traditional) neural networks are trained as discriminative models—usually using cross-entropy or a similar criterion. The deep neural network in a DNN-HMM setup is usually trained to estimate $P(q_t = s|\boldsymbol{x}_t)$, i.e., the probability that the HMM is in state $s \in S$ at time step $t \in \mathbb{N}$, given the observed feature vector $\boldsymbol{x}_t$, cf. Yu and Deng [2015, p. 101]. This means that the deep neural network is applied *frame-wise* on the entire input feature sequence $\boldsymbol{X} := (\boldsymbol{x}_t)_{t=0}^{T}$. Probability $P(q_t = s|\boldsymbol{x}_t)$ is usually estimated independently of the probabilities at previous or future time steps. In particular for feed-forward networks, a fixed window of concatenated input vectors $\tilde{\boldsymbol{x}}_t := (\boldsymbol{x}_n)_{n=t-t_1}^{t+t_2}$ is usually used. Often a symmetric window with $t_1 = t_2 \in \{1, 2, 3, 4, 5\}$ is used as input values for the network. Through the fixed window, the network can consider short time period of input features and not only the current frame to estimate $P(q_t = s|\tilde{\boldsymbol{x}}_t)$. Usually, this makes delta features superfluous.

Bayes' rule has to be applied to obtain observation probabilities from the probabilities estimated by the neural network to utilize the estimation of the neural

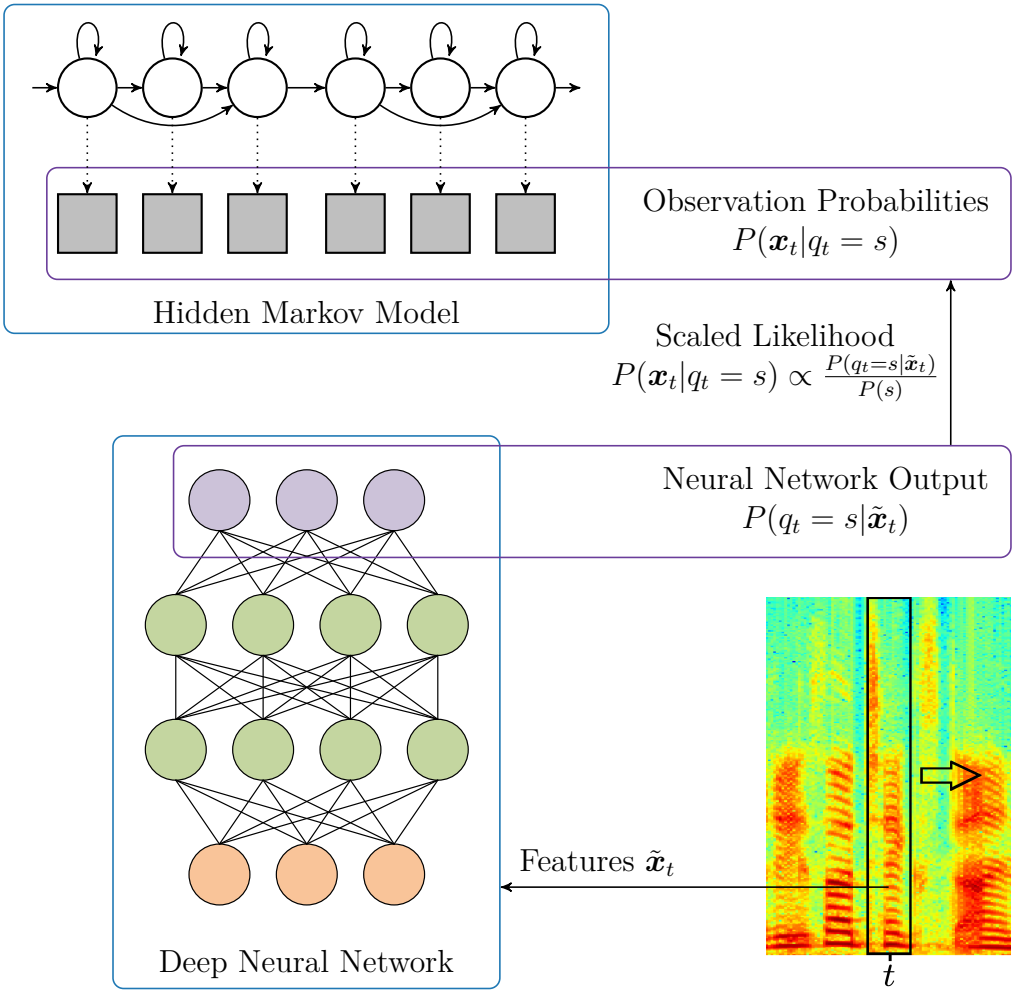


Figure 2.9: Hybrid deep neural network - hidden Markov model speech recognition approach. Spectral features are windowed and used as frame-wise input values for the neural network to estimate HMM state occupation probabilities. These probabilities are then scaled and used as scaled likelihoods to better approximate observation probabilities of the hidden Markov model for recognition.

network in the conventional GMM-HMM setup, i.e.,

$$P(\boldsymbol{x}_t|q_t = s) = \frac{P(q_t = s|\boldsymbol{x}_t) \cdot P(\boldsymbol{x}_t)}{P(s)},$$

cf. Yu and Deng [2015, pp. 101 ff.]. Probability $P(\boldsymbol{x}_t) > 0$ is considered independent from the word sequence and can be ignored during arg max decoding. Thus, the observation probability is approximated with the *scaled likelihood*,

$$P(\boldsymbol{x}_t|q_t = s) \propto \frac{P(q_t = s|\boldsymbol{x}_t)}{P(s)},$$

cf. Morgan and Bourlard [1995]. The number of frames in the training data labeled with state s is counted and divided by the overall number of frames in the training data to estimate probability $P(s)$.

Finally, the HMM acoustic model observation probabilities $P(\boldsymbol{X}|\boldsymbol{s})$ in a hybrid DNN-HMM setup are approximated by

$$P(\boldsymbol{X}|\boldsymbol{s}) = \prod_{t=0}^{T} P(\boldsymbol{x}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}, q_t = s) \approx \prod_{t=0}^{T} P(\boldsymbol{x}_t|q_t = s) \propto \prod_{t=0}^{T} \frac{P(q_t = s|\boldsymbol{x}_t)}{P(s)}$$

cf. Wang et al. [2019]. Overall, the decoding with a hybrid DNN-HMM system is described utilizing the above estimation of $P(\boldsymbol{X}|\boldsymbol{s})$ in the factorized Bayes' decision rule in Equation 2.4 which leads to

$$\hat{\boldsymbol{w}} \approx \arg\max_{\boldsymbol{w} \in \mathcal{W}} \left(P(\boldsymbol{w}) \cdot \max_{\boldsymbol{s} \in \mathcal{S}} \left(P(\boldsymbol{s}|\boldsymbol{w}) \cdot \prod_{t=0}^{T} \frac{P(q_t = s|\boldsymbol{x}_t)}{P(s)} \right) \right).$$

In a hybrid DNN-HMM setup, the hidden Markov model still models temporal relationships through the transition probabilities in $P(\boldsymbol{s}|\boldsymbol{w})$ learned during training of the hidden Markov models. Accordingly, a separate language model still models the probability of word sequences $P(\boldsymbol{w})$. Thus, the hybrid approach preserves the advantages of the GMM-HMM-based approaches described above.

Another advantage of the hybrid approach is that features used as input for the neural network do not have to be statistically independent or decorrelated—as it is often required for GMM-HMMs. Thus, it is possible to use the aforementioned filter bank features, frames of the raw spectrogram, or even entirely other types of features that might be more representative for speech recognition than MFCCs. Additional input features are also often incorporated to enable the model to utilize further knowledge from other models or data. A popular example is the application of speaker-embeddings that enable the system to adapt to different speakers during

recognition. A detailed description of speaker embeddings and speaker adaptation approaches for hybrid models is given later in this section.

Moreover, deep neural networks have a higher potential to learn much better models of data that lie on or near a non-linear manifold [Hinton et al., 2012]. However, there is the disadvantage that the training of hybrid systems becomes more complex and time-consuming. For a hybrid system, a complete GMM-HMM system must be trained first. This system then generates corresponding labels for the training of the deep neural network.

Hybrid systems combine the strong modeling power of deep neural networks with temporal modeling and separate acoustic and language models of GMM-HMM systems. Since the early 2010s, hybrid models outperform GMM-HMMs on various speech recognition benchmarks, sometimes by a large margin [Hinton et al., 2012]. A few years after the emergence of hybrid DNN-HMM systems, this approach almost entirely replaced the conventional GMM-HMM approach in most works.

However, the application of neural networks for acoustic modeling in speech recognition was studied long before the emergence of deep learning. For instance, Bourlard and Morgan [1993] and the survey by Trentin and Gori [2001] describe many approaches that combine hidden Markov models and artificial neural networks acoustic models for continuous speech recognition. Some of these systems even use recurrent neural networks and time delay neural networks, architectures rediscovered for acoustic modeling by recent research and applied in speech recognition systems achieving cutting edge results on many different challenging speech recognition tasks, such as by Peddinti et al. [2015b].

Several aspects limited early works on hybrid systems that used neural networks for acoustic modeling, cf. Yu and Deng [2015, pp. 99–101]. Shallow networks were usually used, and deep neural networks were rarely exploited due to computational limitations. Moreover, the systems often did not utilize context-dependent phones or states but used monophone GMM-HMMs. In early works, the term *CD-DNN-HMM* (context-dependent - deep neural network - hidden Markov models) [Yu and Deng, 2015, p. 101] was often used to distinguish improved hybrid systems that utilized both context-dependent HMMs and deep neural networks from the previous hybrid approaches. Although the previously cited work does not discuss this aspect, it can be assumed that the increasing availability of annotated speech data for training has also greatly contributed to the enormous success of CD-DNN-HMMs in recent years.

2.2.2 Speaker Adaptive Training of Hybrid DNN-HMM Acoustic Models

Speaker adaptive training of hybrid DNN-HMM systems aims to improve the system's recognition quality for unknown speakers that do not appear in the training data—precisely as this is the case with speaker adaptive training of conventional GMM-HMM speech recognition systems. Feature transform approaches using fMLLR transform, which have already been proven beneficial for speaker adaptive training of GMM-HMM systems, have also been studied and improved for neural network acoustic models in recent years, e.g., by Rath et al. [2013]. Such advanced feature space transformations based on fMLLR have been applied—usually as one of multiple speaker adaptive techniques within the system—in different state-of-the-art hybrid speech recognition systems, e.g., by Saon et al. [2017].

However, as discussed above, deep neural networks have greater modeling power and do not have the same constraints to input features as hidden Markov models. This allows for fundamentally different speaker adaptation approaches for hybrid acoustic models that have been proposed and studied in many different works in recent years. Popular examples are approaches that use speaker embeddings—a vector space representation of speakers' voices with a relatively low dimension. Speaker embeddings are popular approaches for *speaker recognition* and *speaker clustering* that allow identifying and verifying speech segments of the same speakers, usually independently from what the speaker said.

The usage of speaker embeddings for speaker adaptive training of hybrid acoustic models has been studied in many different ways. For instance, Xue et al. [2014] proposed and studied three different approaches to use a specific type of speaker embeddings for fast adaptation of hybrid systems: a non-linear feature normalization approach using speaker embeddings, a direct adaptation of the model based on the speaker embeddings, and a joint speaker adaptive training with speaker embeddings. Fast adaptation of hybrid acoustic models to different speakers by performing speaker adaptation through normalization in the feature-space using the well-known speaker embedding *identity vector* (*i-vector*) was studied by Miao et al. [2015b].

The i-vector speaker embedding was proposed by Dehak et al. [2011] for speaker verification. It provides a low-dimensional, text-independent vector space representation of both the speaker and the channel. A main advantage of the i-vector embedding approach is that it can be trained in an unsupervised manner, and no speaker labels are required during training. The prerequisite, however, is that for each speech segment, it is assumed that exactly one speaker is speaking in the recording. The i-vector embedding approach is related to the joint-factor analysis (JFA) [Kenny et al., 2007a,b]. However, in contrast to JFA, which aims to model

disjoint spaces for the speaker and the channel, the i-vector approach models a *total variability space* containing both the speaker and the channel variability.

A relatively simple yet powerful approach for speaker adaptation of hybrid acoustic models has been proposed by Saon et al. [2013]. The authors propose to adapt the acoustic model to the target speaker by supplying the i-vector speaker embedding as input features to the neural network in parallel with the regular acoustic features. Due to its simplicity and increased speaker robustness of models, this approach has prevailed for many systems. It is applied either as a speaker adaptation technique in addition to feature-space transformations such as fMLLR, e.g., in the aforementioned work by Saon et al. [2017], or as the only speaker adaptation approach in the system, e.g., by [Xiong et al., 2018], Peddinti et al. [2015a] and Peddinti et al. [2015b]. The latter two approaches use conventional MFCC features as inputs that are not subject to any normalization. The authors' intention for omitting feature normalization is to allow the network to perform any needed feature normalization itself based on the information encoded in the supplied speaker embedding.

Most approaches using speaker embeddings are not limited to a specific type of embedding. Other speaker embeddings, such as *x-vectors*, proposed by Snyder et al. [2018], are assumed to be suitable for speaker adaptation in ASR as well, cf. Raj et al. [2019]. Moreover, additional embeddings, e.g., for channel characteristics, such as *r-vectors*, proposed by Khokhlov et al. [2019], that model room acoustics might also be beneficial to improve adaptation to unseen conditions.

2.3 End-to-End Speech Recognition

In recent years, deep learning systems achieved impressive results in many different machine learning tasks. Initially, deep neural networks were often used as one component in a larger processing pipeline, for example, with feature extraction and other models. Usually, such pipelines rely on expert human knowledge—for example, to design an appropriate feature extractor or phonetic pronunciation modeling. In recent years, however, there has been a strong trend towards *end-to-end learning* that aims at training a single model that replaces the entire processing pipeline. Such systems require no or very little human expert knowledge and learn all the necessary processing steps on their own.

Automatic speech recognition based on DNN-HMMs is one example of a deep neural network representing one component in a rather complex processing pipeline. *End-to-end speech recognition* aims at simplifying this pipeline for training and application of speech recognition systems by removing the whole hidden Markov model aspect from the pipeline. It became a field with enormous research interest since the late 2010s.

A common definition of *end-to-end* in the field of speech recognition does not seem to exist. Two dominant definitions of end-to-end speech recognition can be found in the literature:

- Some authors, such as Miao et al. [2015a], Watanabe et al. [2017], and Hadian et al. [2018], define end-to-end models as single DNNs that do not rely on previously trained models—such as HMMs. Thus, these end-to-end models are trained in one stage, cf. Hadian et al. [2018]. In contrast to bootstrap training, they do not require forced alignment of training data or state-tying. These end-to-end models map audio feature sequences, such as MFCC or FBANK features, to sequences of phonetic units—such as phones, biphones, and triphones—or grapheme units. Therefore, additional models based on human knowledge might still be required, e.g., in the form of phonetic pronunciation lexicons.

- Other authors exhibit a stricter notion of end-to-end in speech recognition that is more in line with the general understanding of end-to-end in deep learning. They define end-to-end models as DNNs that map an audio sequence directly to a sequence of characters or words. This definition is found, among others, in the works by Graves and Jaitly [2014], Prabhavalkar et al. [2017], Zeyer et al. [2018], Wang et al. [2019], and Andrusenko et al. [2020]. Since these models do not use phonetic units, a pronunciation lexicon is not required, and decoding is highly simplified.

Except for the choice of output, the two definitions coincide. As described later in this section, end-to-end speech recognition is not a relevant approach for the presented work. Thus, there is no need to distinguish between these definitions for the further course of the work.

In conventional, hybrid ASR systems, the neural network is applied frame-wise on the input sequence. Thus, the neural network's input and output values always are of fixed, known dimensions. However, in end-to-end speech recognition, a sequence-to-sequence mapping is required. Sequences pose a multitude of challenges for deep neural networks, cf. Sutskever et al. [2014]. One main challenge in sequence-to-sequence modeling is that the lengths of the input and output sequences are different by a substantial margin and a-priori unknown.

2.3.1 End-to-End ASR Approaches

In end-to-end speech recognition, the input sequence is a sequence of audio features such as MFCCs or the raw audio signal. The desired output is a sequence of recognition results, such as letters, entire words, or sub-word units. The expected

output sequence is usually much shorter than the sequence of input features. End-to-end speech recognition can be divided into different categories based on the approach utilized to overcome the challenge of sequence-to-sequence modeling:

- *Connectionist temporal classification* (*CTC*) was proposed by Graves et al. [2006] as a training method for sequence labeling of unsegmented data with recurrent neural networks. CTC transforms the neural network outputs to conditional probability distributions over label sequences.

- *RNN-transducers* were introduced by Graves [2012] to resolve the conditional independence assumptions of CTC. The RNN-transducer model extends the CTC approach by defining a distribution over output sequences of all lengths and by jointly modeling both input- and output-sequence dependencies. The RNN-transducer architecture utilizes an additional subnetwork that recurrently models output labels.

- *Attention-based models* represent the third main direction of sequence modeling for end-to-end speech recognition. These models are encoder-decoder sequence models, cf. Sutskever et al. [2014], that utilize the well-known *attention*-mechanism for automatic speech recognition. Several variants to the attention mechanism have been proposed in recent years to improve various aspects for diverse tasks. The first popular attention mechanisms have been proposed by Bahdanau et al. [2015] and Luong et al. [2015]. A prominent attention model variant used in many current deep learning applications is the *transformer* proposed by Vaswani et al. [2017]. Different works studied attention-based models for speech recognition, such as Bahdanau et al. [2016], and the prominent *listen, attend, and spell* model proposed by Chan et al. [2016].

Different works aim at combining attention models with CTC or RNN-transducer models to take advantage of both approaches in one system. A combination of attention with CTC is proposed by Watanabe et al. [2017]. Combinations of attention with RNN-transducers (*transformer attention-based encoder-decoder*) are referenced in the comparison work on end-to-end systems by Li et al. [2020].

2.3.2 End-to-End LF-MMI

Another approach proposed for end-to-end speech recognition is the end-to-end variant of the well-known *lattice-free maximum mutual information* (*LF-MMI*) approach. LF-MMI, in general, is an extension proposed by Povey et al. [2016] to the hybrid DNN-HMM approach. LF-MMI aims to utilize the advantages of end-to-end speech recognition research for hybrid models. Due to great performance

on many different speech recognition benchmarks, it quickly became state of the art in speech recognition. Therefore, it is discussed in detail in Section 2.4.3.

The end-to-end implementation of LF-MMI was proposed by Hadian et al. [2018]. However, it depends on the definition of *end-to-end* whether this approach is indeed *end-to-end speech recognition*. End-to-end LF-MMI is not a single neural network model, as the aforementioned approaches are, but it utilizes hidden Markov models to model the temporal sequences. However, these hidden Markov models are not trained in a prior stage, as with hybrid models, but the entire system is trained in one step. Also, no state tying is performed in end-to-end LF-MMI.

The end-to-end variant is very similar to the original LF-MMI approach but does not achieve as promising speech recognition results. For the sake of completeness, this approach is listed in this section on end-to-end speech recognition. However, due to the above reasons, it is not considered a relevant approach for the present work.

2.3.3 Relevance of End-to-End Speech Recognition for the Present Work

End-to-end speech recognition approaches greatly simplify the overall speech recognition problem and allow the development of innovative, promising systems. However, despite all advantages that end-to-end speech recognition promises, in the years 2017–2020, there are major drawbacks compared to conventional hybrid systems. Many works show that end-to-end systems' performance is worse compared to hybrid systems, and only some works show comparable or slightly better performance of end-to-end systems, cf. the overview given by Wang et al. [2019]. This was the case for most works, especially when the present work started in early 2017—but still prevails towards the end of the present work. Even though the performance gap is constantly shrinking, end-to-end only slowly surpass conventional systems.

Different works suggest that the performance gap could highly depend on the amount of training data. For instance, Lüscher et al. [2019] performed experiments on the widely used Librispeech corpus [Panayotov et al., 2015] to compare end-to-end and hybrid systems. The authors proposed a hybrid and an end-to-end model that both achieved the best result on Librispeech published in the literature to the current time for the respective category.

The authors showed that the margin between the end-to-end and hybrid model was significantly larger when a reduced 100-hours subset from the total 960-hour training data was used. Furthermore, the authors observed that the gap between the conventional hybrid and the end-to-end system increases from 15 % to 40 %

when evaluating the model on the *other* test set portion of Librispeech that contains more challenging speech instead of the *clean* set. This could indicate low robustness of end-to-end models or susceptibility to deviations between training and test data. Similarly, Andrusenko et al. [2020] state that end-to-end speech recognition systems face problems in noisy, far-field, and low-resourced conditions.

As discussed in Chapter 3, the German oral history interviews that are the primary focus of this work pose all and more than the previously mentioned challenges: low-resourced training data, challenging speech, noisy recording conditions, and reverberation. The German oral history data are so challenging that the baseline off-the-shelf speech recognition system showed word error rates above 50 % during the start of the present work. Therefore, after the literature review, the decision was not to include end-to-end systems in the research work and focus on hybrid models. The previously cited works on end-to-end models published in the course of the present work indicate that these challenges of end-to-end modeling remain, and thereby a posteriori justify the decision made in 2017. However, with the ever-increasing accuracy of end-to-end models in the early 2020s and the greater simplicity and flexibility, this decision may need to be reconsidered in future works.

2.4 Sequence Discriminative Training of Neural Networks

Concurrent with research on end-to-end speech recognition systems, research on hybrid models also continued and yielded promising new approaches. Numerous works proposed and studied methods to further improve neural network training for hybrid systems. One notable direction is *sequence discriminative* training, studied, among others, by Kingsbury [2009] and Wang and Sim [2011] for hybrid systems. Similar to end-to-end training, these studies aim at optimizing the neural network with sequence-level objectives instead of frame-wise cross-entropy. However, in contrast to end-to-end systems, sequence discriminative trained hybrid systems still rely on temporal sequence modeling with hidden Markov models and bootstrapped training—usually with context-dependent phones.

2.4.1 Sequence Discriminative Training Criteria

Sequence discriminative training is well-known from conventional GMM-HMM-based speech recognition systems, cf. Vertanen [2005], and has been proven to significantly increase recognition accuracy in many scenarios. Different criteria are used in sequence discriminative training. The most common approaches can be categorized as *maximum mutual information* or *minimum Bayes risk*, cf. Veselý

et al. [2013] and Yu and Deng [2015, pp. 137 ff.]. A summary of these criteria is given in the following.

Maximum Mutual Information

Maximum mutual information (MMI) training proposed by Bahl et al. [1986] aims at maximizing the mutual information between an acoustic observation sequence and the corresponding word sequence. For a training set with $N_U \in \mathbb{N}$ samples, with $\boldsymbol{X}_u := (\boldsymbol{x}_{u,k})_{k=1}^{T_u}$ as the sequence of observed features and $\boldsymbol{s}_u$ as the sequence of states, both corresponding to training utterance u, the MMI criterion is defined as:

$$J_{\text{MMI}} := \sum_{u=1}^{N_U} \log \left(\frac{P(\boldsymbol{X}_u|\boldsymbol{s}_u, \boldsymbol{\theta})^\kappa P(\boldsymbol{w}_u)}{\sum_{\hat{\boldsymbol{w}} \in W} P(\boldsymbol{X}_u|\boldsymbol{s}_{\hat{\boldsymbol{w}}}, \boldsymbol{\theta})^\kappa P(\hat{\boldsymbol{w}})} \right), \tag{2.6}$$

where $\kappa \in \mathbb{R}$ is an acoustic model scaling factor and $P(\boldsymbol{w})$ is the probability of a word sequence, cf. Veselý et al. [2013].

The numerator in Equation 2.6 is the likelihood of the sequence of correct states for the training utterance to create the observed sequence of features. The denominator is the sum of likelihoods for all possible sequences of states to produce the observed feature sequence. Optimizing the system with the MMI criterion maximizes the numerator and, at the same time, minimizes the denominator. Thus, for a given sequence of features, the MMI criterion attempts to make the correct word sequence more probable and all incorrect word sequences less probable, cf. Vertanen [2005]. A modified variant of MMI called *boosted MMI* was proposed by Povey et al. [2008] to improve recognition results by boosting the likelihood for state sequences in the denominator with higher phone errors.

Minimum Bayes Risk / Minimum Phone Error

Minimum Bayes risk (MBR) criteria [Kaiser et al., 2000, 2002] aim at minimizing the risk of the wrong classification on the training data. MBR is a family of criteria with variants for different granularity of labels that have been studied for hidden Markov models, cf. Yu and Deng [2015, pp. 140 ff.]. MBR criteria are defined by a loss function of the following type:

$$J_{\text{MBR}} := \sum_{u=1}^{N_U} \left(\frac{\sum_{\boldsymbol{w} \in W} P(\boldsymbol{X}_u|\boldsymbol{s}_{\boldsymbol{w}}, \boldsymbol{\theta})^\kappa P(\boldsymbol{w}) A(\boldsymbol{w}, \boldsymbol{w}_u)}{\sum_{\hat{\boldsymbol{w}} \in W} P(\boldsymbol{X}_u|\boldsymbol{s}_{\hat{\boldsymbol{w}}}, \boldsymbol{\theta})^\kappa P(\hat{\boldsymbol{w}})} \right) \tag{2.7}$$

where $A(\boldsymbol{w}, \boldsymbol{w}_u)$ is referred to as *raw accuracy* and depends on the label to be optimized—e.g., words, phones, or states. An overview of commonly used definitions of $A(\boldsymbol{w}, \boldsymbol{w}_u)$ for different objectives is studied by Povey and Kingsbury [2007]

for GMM-HMM systems. *State-level minimum Bayes risk* (sMBR) became popular for cutting-edge HMM-based acoustic models where the definition of $A(\boldsymbol{w}, \boldsymbol{w}_u)$ is the total number of correct states in the label sequence.

Since the denominator of MBR and MMI are equal, both criteria minimize wrong classifications. However, the numerator of MBR in Equation 2.7, sums different likelihoods and weights them with the raw accuracy $A(\boldsymbol{w}, \boldsymbol{w}_u)$ depending on their influence on the expected resulting classification accuracy. In simple cases, for sMBR, the raw accuracy can be defined as the sum of correct state labels.

The well-known criterion *minimum phone error* (MPE) was proposed by Povey and Woodland [2002] simultaneously to MBR. This criterion is a smoothed approximation for the phone error rate of the model. Since it uses the same general Equation 2.7 but applies it to phones, it can be considered a member of the MBR family, cf. Yu and Deng [2015, pp. 140 ff.]. MPE uses a raw phone accuracy for $A(\boldsymbol{w}, \boldsymbol{w}_u)$, i.e., the number of correct phones, to estimate and minimize the resulting phone error of the system.

2.4.2 Sequence Discriminative Training of Neural Network Acoustic Models

Recent works on hybrid DNN-HMM systems studied applying the same sequence objectives from hidden Markov model systems to acoustic model neural network training. For example, Veselý et al. [2013] studied and compared sequence discriminative training with MMI, MPE, and sMBR criteria to train a feedforward deep neural network for a hybrid system. The sequence discriminative trained models all outperformed the cross-entropy trained baseline. These models achieved state-of-the-art performance at that time on the 300-hour Switchboard conversational telephone speech task [Godfrey et al., 1992]. Two years later, Voigtlaender et al. [2015] studied the application of sequence discriminative training with the MMI criterion for bidirectional LSTM-RNNs. The authors state that bidirectional LSTMs utilize the whole sequence to incorporate more contextual information than feed-forward neural networks. With their experiments, Voigtlaender et al. showed that sequence-discriminative training gives substantial improvements over a cross-entropy trained LSTM baseline.

Traditionally, sequence discriminative training of neural networks utilizes the same *lattice-based* approaches that have been developed for hidden Markov model training, cf. Kingsbury [2009]. A cross-entropy system is first trained to generate the word lattices for sequence discriminative training with a weak language model, cf. Povey et al. [2016]. These lattices are then used to approximate all possible word sequences used in sequence discriminative objective functions. Then, the

cross-entropy trained model is used as initialization to train the neural network acoustic model using the sequence discriminative criterion.

2.4.3 Purely Sequence-Trained Neural Networks Acoustic Models Using Lattice-Free MMI

One of the publications with the biggest impact on sequence discriminative training of hybrid acoustic models in recent years is the work by Povey et al. [2016]. In their work, Povey et al. successfully apply concepts from CTC training to sequence discriminative neural network training with MMI since both criteria generally aim at optimizing the conditional likelihood of the correct transcript.

Povey et al. propose a neural network acoustic model training with MMI without initialization of the model with a cross-entropy system—as it is traditionally done for sequence discriminative training. The authors propose a *lattice-free* implementation for the numerator and denominator likelihoods from the MMI criterion to train the system directly on GPUs. Instead of storing all possible label sequences in a huge lattice for the input feature sequence of each training utterance—where each path through the lattice represents one label sequence and its corresponding likelihood—the lattice-free implementation represents the sequences as a graph. It aims at making the graph as small as possible so that it fits in the GPU. This means that the graph for the denominator is the same for all utterances. Since GPU-based computations benefit from synchronized memory access, the sum of the denominator can be calculated efficiently in the GPU for each training utterance.

To reduce the size of the graphs, lattice-free MMI (*LF-MMI*) utilizes a phone-level language model instead of a word-level model for the MMI criterion. Moreover, to reduce space and time complexity during training—and also improve efficiency in decoding—*LF-MMI* models use a 3-fold reduced (*subsampled*) frame rate and accordingly simplified HMM topology with constant, evenly distributed transition probabilities.

Purely sequence-trained systems tend to overfit and show poor generalization to unseen data, cf. Vertanen [2005]. LF-MMI training utilizes different regularization techniques to overcome this limitation. These are L2-norm output regularization, a *cross-entropy regularization*—realized as multi-task learning of LF-MMI and cross-entropy with two output layers—and *leaky* hidden Markov model transition probabilities. More implementation details are different in LF-MMI from traditional MMI, such as fixed-length chunks for training. These are described in detail in the original work [Povey et al., 2016] for further reading. In some literature,

LF-MMI systems are sometimes referred to as *chain* models due to the naming of LF-MMI in the Kaldi implementation.[4]

Povey et al. showed that LF-MMI outperforms cross-entropy and sMBR trained hybrid systems for many model architectures and various data sets. Numerous subsequent works utilized LF-MMI to achieve cutting-edge results on different challenges at their respective time. To name but a few examples in the following that have been published during the course of the present work:

- The best-performing systems of the *2018 5th CHiME Speech Separation and Recognition Challenge (CHiME-5)* [Barker et al., 2018], e.g., [Du et al., 2018], [Kanda et al., 2018], utilize LF-MMI.
- Furthermore, all *ASRU 2019 Mandarin-English Code-Switching Speech Recognition Challenge* participants used LF-MMI for the hybrid ASR system track, cf. Shi et al. [2020].
- The two speech transcription systems of the *MGB-5 Recognition and Dialect Identification of Dialectal Arabic Speech Challenge* [Ali et al., 2019] outperform the baseline—that used LF-MMI itself—also used LF-MMI training in some way, cf. Ahmed et al. [2019], [Khurana et al., 2019].

These different challenges present extremely difficult scenarios for automatic speech recognition systems in our current time. The widespread application of LF-MMI demonstrates its capabilities in these challenging use cases.

2.5 Summary

Our thorough presentation of the major pioneering works in automatic speech recognition in this chapter indicates that automatic speech recognition is a field with enormous research efforts in various subfields. Besides the fundamental aspects of speech recognition, in our provided overview, we focused on pioneering modeling approaches that have prevailed until now and are applied in cutting-edge systems. In particular, we focused on acoustic modeling techniques since these are the main research focus of the present work. Our choice of focus means that other significant work, such as advances in language modeling, has gone unmentioned. However, this does not mean that these are less significant for speech recognition per se but do not correspond to the focus of the present work.

In this chapter, we first explained the theory of hidden Markov models that have been the core component of powerful speech recognition systems for many

[4] An overview of Kaldi and other automatic speech recognition toolkits can be found in the appendix in Section A.1.

decades. We presented how the general speech recognition decoding problem can be reformulated using Bayes' decision rule. Three different models—the acoustic model, the phonetic pronunciation lexicon, and the language model—can be used independently to address different aspects of the problem formulation.

Building on the presentation of hidden Markov model speech recognition systems, we have described the application of deep neural networks in hybrid models that combine hidden Markov models and neural networks for acoustic modeling. Over the past decade, these hybrid models have vastly improved acoustic modeling for automatic speech recognition. We presented commonly used feature extraction methods, feature space transformations, and speaker adaptive training approaches for conventional hidden Markov models and hybrid systems.

With our description of end-to-end automatic speech recognition systems, we credit this emerging and highly promising research strand that has been gaining enormous research interest in recent years. We presented the three most common fundamental end-to-end speech recognition approaches and how they significantly simplify automatic speech recognition by removing the hidden Markov models from the processing pipeline. We discussed why end-to-end approaches were not considered to be studied for the oral history challenge in the presented research work.

Furthermore, in this chapter, we presented research efforts in sequence discriminative training of neural network acoustic models. To some extent, this research shows some parallels to end-to-end training since it attempts to perform sequence-based optimizations—but on hybrid models that utilize hidden Markov models for temporal modeling instead of sequence-to-sequence neural networks. The recent summit of this research strand is LF-MMI model training that enables hybrid DNN-HMM systems to achieve cutting-edge results for various speech recognition challenges and use cases.

Despite all advantages of deep neural networks in automatic speech recognition, challenges remain open that are discussed in the following main chapters of the present work. Usually, deep neural networks require lots of training data to achieve robust generalization and fully unfold their modeling power in real-world applications. This is particularly the case for challenging automatic speech recognition tasks. Usually, large amounts of annotated speech are required. Such data often is not available for the desired use case. A major question to be answered in the following chapters is how deep neural networks can be exploited in real-world applications when only little training data is available for a concrete use case.

3 Automatic Transcription of Oral History Interviews

In the humanities, *oral history* refers to conducting and analyzing interviews with contemporary witnesses to historical events. For oral history researchers, transcription is essential to find interviews of interest for the respective research question in large archives and to find and examine relevant passages within very long interviews. Currently, transcribing, labeling, and annotating these speech recordings is often performed entirely manually.

Speech recognition systems offer a promising opportunity to automatically process oral history archives to reduce the resources needed to transcribe these interviews significantly. However, the past has shown that applying off-the-shelf speech recognition technology on oral history interviews usually results in poor recognition performance. Compared to other speech recordings, oral history interviews pose several challenges for speech recognition systems. Often the quality of the audio recording is low due to different distortions occurring simultaneously, e.g., background noises, poor recording equipment, room reverberation, and a large distance between the speaker and the microphone. Furthermore, the interviewed persons usually are elderly persons speaking very spontaneously. To overcome these limitations and make speech recognition applicable to oral history interviews, the Fraunhofer IAIS has worked on adapting the automatic transcription system *Audio Mining* to oral history interviews in a research project since 2015. This chapter presents and discusses the automatic transcription of oral history interviews. We discuss the challenges that these interviews pose for existing systems and conduct preliminary experiments to identify the aspects of systems that need improvement.

This chapter is structured as follows. We first introduce the research project KA^3 that funded major parts of the presented research work, the oral history use case, and archives considered in the project in Section 3.2. Furthermore, we provide a systematic review of related works investigating automatic speech recognition for oral history interviews for different languages in Section 3.3. With this review, we expose the existing research gap in this field and give an overview of the challenges of oral history interviews for speech recognition in general and particularly for the data archive we study. In Section 3.4, we present a representative German oral history test set developed and proposed during the present research work, which is used to investigate the performance of proposed systems and approaches. In this

context, we give an overview of the other training and test data used in this work. As a foundation for investigating different approaches in the following chapters, we perform a preliminary investigation of the influence of different language models on the oral history ASR task in Section 3.5. The chapter concludes in Section 3.6 with an investigation of the estimated *human word error rate* for German oral history interviews. We show that word-accurate transcriptions of oral history interviews are challenging even for humans. The estimated human word error rate can be understood as a lower bound of the expected performance of speech recognition systems on such interviews. In Section 3.7, we summarize the chapter's findings and contributions.

3.1 Thesis Author Contribution

Sections of this chapter are covered in the publications:

> Michael Gref, Joachim Köhler, and Almut Leh. Improved transcription and indexing of oral history interviews for digital humanities research. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 3124–3131. European Language Resources Association (ELRA), 2018a. URL https://aclanthology.org/L18-1493

> Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*, 2022b. URL https://arxiv.org/abs/2201.06841

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author. The oral history data sets described in Section 3.4.6 and Section 3.4.10 were provided by co-authoring project partners in coordination with the thesis author. Respective author contributions are given in Appendix C.

3.2 The Oral History Use Case

3.2.1 The KA³ Project

Major parts of the present research work were carried out as a part of the project *KA³ - Kölner Zentrum für Analyse und Archivierung audiovisueller Daten* (*Cologne Center for Analysis and Archiving of Audiovisual Data*) funded by the *German Federal Ministry of Education and Research* between 2015 and 2020. The overall

goal of the KA^3 project was to establish a digital center to make audiovisual data sustainably usable for researchers from the humanities and cultural sciences. In addition to archiving, a key goal of the project was to develop, adapt, and apply AI-based methods for automatic indexing and searching of documents as well as video and audio recordings.

In the project's scope, cf. Leh et al. [2018], speech data collections from two different research fields have been selected as use cases to study: The *oral history* and the *interaction* scenario. The interaction scenario represents the application of speech processing technologies in the field of linguistics. Here, mainly human interaction processes between two persons are analyzed. The studied objects are speech recordings in which two persons converse naturally, and *backchannels* of the person currently listing to the speaker frequently occur—such as *uhm*, *okay*, and *uh-huh*.

The research of this thesis began in 2017 in the KA^3 project. It focused on developing automatic speech recognition training methods to make this technology reasonably usable for oral history interviews. For the research work, close cooperation has been established between the Fraunhofer IAIS and the archive *Deutsches Gedächtnis* (*German Memory*) in the *Institute for History and Biography*, University of Hagen in Germany.

Due to different requirements of both the oral history and the interaction use cases, and partly dissimilar challenges in both scenarios, the interaction scenario was only considered in passing in this research work. A major challenge in the interaction scenario was the lack of suitable test data for evaluation at the beginning of the research work. When these were available in later stages, we also examined subsequent experiments on interaction test data. However, automatic transcription of data from the interaction use case utilizing speech recognition had only a secondary interest for the linguistics researchers since backchannel and interaction patterns between the speakers are central to their investigation. For the presented research, we use this and several other German ASR evaluation sets from different domains as a *control group* to examine the real-world performance of the proposed system for unseen data and to verify that improvements on oral history data do not negatively impact the overall recognition performance on other real-world data.

3.2.2 Oral History Interviews and Archives as Sources for the Humanities

Historical 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. As described by the historian Dr. Almut Leh, head of the archive *Deutsches Gedächtnis*, in our joint publica-

tion [Gref et al., 2018a], biographical research emerged in almost all areas of the humanities since the early 1980s: sociology and pedagogy, ethnography and ethnology, historical and literary studies as well as psychoanalysis and psychology. Research conducting and analyzing interviews with contemporary witnesses has become known as *oral history* in the historical sciences. In Germany, this research was mainly focused on 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 presenting historical information in exhibitions, documentations, and movies without using witness accounts to the relevant events. This method in which most of the interviews in question were conducted is characterized by the interviewer encouraging the interviewee to freely narrate their life story rather than structuring the interview around questions. 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 represents a highly individual testimony. The interviewee has presented large parts of their life story and worldview that are often unguarded and sometimes contradictory. Due to the open character of the narration and the life-story dimension, such an interview is valuable for more than a single analysis, cf. [Köhler et al., 2019; Leh et al., 2019]. It can be explored for different topics, even more so since many witnesses have died in the meantime, leaving only their recorded accounts. For the same reason analyzing and archiving oral history interviews is valuable and challenging.

Today archives, museums, websites, and documentation centers preserve and provide oral history interviews for historical research, social sciences, and other humanities. Since the beginnings of oral history, many university and non-university projects have been initiated, in which interviews with contemporary witnesses were collected and evaluated. In Germany, for instance, since 2006, the *Center for Digital Systems* (*CeDiS*) of Free University of Berlin, [Pagenstecher, 2019a,b], provides access to several major oral history archives focusing on the Second World War and the period of National Socialism in Germany. These are the *Visual History Archive* of the *USC Shoah Foundation*, the *Fortunoff Video Archive* of the Yale University, the *Forced Labor 1939–1945* interview archive (introduced by Leh and Tausendfreund [2017]), the British-Jewish collection *Refugee Voices*, the interview archive *Memories of the Occupation* in Greece, and the aforementioned archive *Deutsches Gedächtnis* of University of Hagen.

Regarding the importance of transcription, Pagenstecher states that indexation and full-text search make the long recordings in archives accessible via CeDiS. However, this requires a huge effort in manual transcriptions. Although automatic speech recognition systems have made significant progress in recent years, the

transcription quality of such systems on oral history interviews is usually low. This limits the usability of automatically generated transcriptions for research—especially considering the high standard expectations of the oral history research community.

The *Stiftung Haus der Geschichte der Bundesrepublik Deutschland* (*HdG*; House of the History of the Federal Republic of Germany Foundation) is another important institution in the field of German oral history. With the *Zeitzeugenportal*[1], the foundation hosts an internet platform accessible to the general public that contains a large collection of oral history interviews on several topics of German history in the 20th and 21st century. On the Zeitzeugenportal, more than 8,000 clips from around 1,000 interviews can already be found. The large database offers interesting historical content and allows the viewer to empathize with what they have experienced through emotionally charged stories. This online service gives users the chance to hear about people's stories at any time.

3.2.3 The Oral History Archive "Deutsches Gedächtnis"

Leh [2018] states that the archive *Deutsches Gedächtnis* (German Memory) was founded in 1993 and provides about 3000 oral history interviews conducted from the late 1970s to this day in more than 100 projects using various recording technologies and interview settings.

The average length of the interviews is approximately 3.5 hours. As described above, the interview usually is not structured by questions. Instead, it is open for the course of memories coming into the interviewee's mind when telling their entire life story from birth and childhood into the present. The interviewees present in the archive were born between 1895 and 1980, cf. [Gref et al., 2018a]. The original analog recordings of the 2,000 audio and 500 video interviews have been digitized. Although retrieval and analysis are based on transcription, only half of the interviews had been transcribed at the start of the presented research work in 2017. At that time, only ten percent of the transcripts were time aligned so that the transcript could be used for subtitling or for jumping to the position of a particular word in the audio or video recording. All interviews are equipped with archival, technical, and biographical metadata. Due to the interviewees' privacy, not all of these interviews are available for technology research. The data used in this research is described in detail in Section 3.4.

[1] https://www.zeitzeugen-portal.de

3.3 Systematic Review of Challenges Oral History Interviews Pose for ASR

Automatic transcription systems, such as the Fraunhofer IAIS Audio Mining system[2], offer great advantages in analyzing and archiving interviews for oral history archives. Speaker-diarization, time-aligned transcription, and indexing with keywords are essential for retrieving interviews in large archives and sections within long interviews fast and reliably. Before the advent of automatic transcription systems, all of these steps had been performed manually. The huge effort in time and human resources required to do this severely limits the utilization of oral history interviews for digital humanities research.

The Audio Mining system allows historians to process vast amounts of oral history recordings and improve their research based on these interviews. This is achieved by transcribing the recordings automatically and providing additional speech analysis results in a convenient interface to structure the content. Regarding archiving and retrieval, the Audio Mining system allows full-text search with direct access to the audiovisual media file.

Automatic transcription systems such as Audio Mining facilitate new oral history research approaches in quantity and quality. Comparative studies and quantitative analysis become feasible by covering more data without an enormous manual effort. Furthermore, speech technologies allow analyzing verbal and non-verbal aspects of communication more deeply, thus opening new dimensions for qualitative research, cf. [Köhler et al., 2019; Leh et al., 2019].

Since the early days of oral history, oral historians have insisted on the oral nature of their sources, cf. [Gref et al., 2018a]. Following this demand, the interview recording is the primary source and should be the main subject of research. In consequence, the transcript is only a necessary additive for the analysis. However, in current practice, the transcript is often the only source for interpretation and analysis, replacing the audio completely. Subtitled audiovisual recordings overcome this discrepancy. Thus, they fulfill the essential demand to analyze not only the transcript but the entire oral source. Automatic transcription systems, such as Audio Mining, allow taking full advantage of the untapped potential of oral history leading back to its original roots, cf. [Köhler et al., 2019; Leh et al., 2019].

However, all these advantages are only given if a sufficiently good transcription quality can be achieved. Attempts to use transcription systems for oral history are not new. For two decades, attempts have been made to use speech recognition for this purpose. Until today, there seems to be no satisfactory solution. This

[2] An overview of this system can be found in the appendix in Section A.2. The models trained and improved in the presented research work were integrated into the Audio Mining system for real-world application.

is due to the extraordinary challenges that oral history interviews pose to speech recognition systems.

This section reviews related works that investigated and improved speech recognition for automatic transcription of oral history interviews. Based on these related works, we present a non-statistical, systematic review of the challenges that oral history interviews pose for speech recognition systems—and put them into the context of the German data examined in this work, highlighting the research gap in this field.

3.3.1 Related Work in the Field of ASR for Oral History

The social relevance of preserving the experienced memory of the gradually dwindling generation of contemporary witnesses of the Second World War for future generations is probably quite undisputed. Automatic speech recognition can play a significant role in making the work of historical researchers significantly easier and in making large, previously untapped archives of interviews available for further research. However, the research interest of computer science for the oral history ASR use case seems to be quite limited so far. This is shown by the comparatively low number of publications covering this topic and the relatively low citation count of published papers—especially compared to the enormous number of publications for other ASR challenges in the last decade. In the following, we provide an overview of selected research work that aims at studying and improving automatic speech recognition for oral history interviews in different languages.

The application of automatic speech recognition technology to transcribe and index oral history interviews has started with the MALACH (multilingual access to large audioarchives) project [Psutka et al., 2002], where the interviews of the *Survivors of the Shoah Visual History Foundation* (VHF) were processed with state-of-the-art speech recognition in 2002. The activity's main challenge was the variety and quality of the recordings and the variety of the languages.

Early works on training ASR systems for English data of MALACH by Ramabhadran et al. [2003] achieved a word error rate (cf. Section 3.4.13) of 43.8 % with HMM systems using fMLLR speaker adaptive training and the combination of training data from MALACH and the Switchboard corpus. Siohan et al. [2004] investigated possible speech recognition error sources for this data. The authors analyzed the segmental signal-to-noise ratio (SNR) on an English 65 hour subset, the syllable rate on a 200-hour subset, and the age distribution of the speakers. For their ASR system, the authors found the syllable rate and SNR to be the most dominant factor on the word error rate. The average SNR on the used subset is 23 dB with a long-tailed distribution towards lower SNRs and short-tailed towards higher SNRs. Thus, in our interpretation of the reported results, it can be considered relatively clean or slightly noisy for most of the data, with some exceptions

containing strong distortions. The authors report that noise compensation in the log-spectral domain on the test data results in only a slight improvement of 1.1 % absolute.

Psutka et al. [2005] continued research on the MALACH corpus, developing speech recognition systems for Czech, Russian, and Slovak oral history interviews. The authors state that accented and spontaneous speech characterize these interviews. Particularly for Russian, various regional pronunciation variants were a challenge. The age of the speakers also affected the quality and fluency of the speech. Using 84–100 hours of training data for each language, the proposed GMM-HMM systems achieved 34.5–45.8 % word error rates. Byrne et al. [2004] also investigated speech recognition for Czech and English MALACH interviews. Using 65–84 hours of annotated interviews, the authors achieved a word error rate near 40 % for both languages. The authors list age-related changes in the way of speaking, highly variable speaking rates, and heavily accented speech as challenges for machine and human transcribers. Mihajlik et al. [2007] investigated Hungarian oral history interviews of the MALACH corpus and achieved a 37.5–41.1 % word error rate using unsupervised speaker adaptation on GMM-HMM systems.

Oard [2012] described how speech recognition technology can be used for oral history research and summarized their research efforts for recordings from MALACH. Oard concludes that in 2012 fully automatic transcription on the challenging content of the MALACH could not be achieved yet and is not yet even on the horizon.

Hessen et al. [2013] describe the use of speech recognition to transcribe Dutch oral history archives. The authors state the word error rate is above 40 % for Dutch oral history interviews at the time of publishing. Automatic speech recognition with such an error rate does not produce results that meet the requirements of automated subtitling or transcription of recorded interviews—but could be used for indexing and searching in large archives. Hessen et al. further present three reasons for the poor recognition performance on their oral history data:

1. Poor recording quality of interviews due to unsuitable conditions during recording. This might occur, for instance, since the interviewers often have to improvise when recording the interview with the eyewitness.

2. The interviewed persons usually do not speak fluently nor grammatically correct.[3]

3. The interviewees have dialects or are non-native speakers that have an accented way of speaking.

[3] This is often a problem of spontaneous speech that we observe in automatic speech recognition and it is not limited to oral history interviews—but may occur more intensely there due to the open character of the interview.

Salesky et al. [2016] studied *keyword ASR* for English oral history data. The authors worked with oral history data collected by *StoryCorps*, an American non-profit organization that provides volunteer participants the opportunity to record, share and preserve their life stories. The speech recognition results are used, among others, to assess and compare a *human search capability* on oral history data. The authors perform experiments with 100 hours of *in-domain* training data and with *out-of-domain* training data from the Wall Street Journal (WSJ) corpus [Paul and Baker, 1992] and Fisher corpus [Cieri et al., 2004]. For evaluation, the authors use a 50-hour in-domain evaluation set. A GMM-HMM acoustic model approach trained with HTK was used for the experiments.

Salesky et al. achieved a 38.5 % word error rate on the 50-hour evaluation set with in-domain training data. Experiments with only out-of-domain training data for the acoustic and language model yield worse word error rates: 68.1 % with WSJ and 49.5 % with Fisher. The authors also performed experiments on different combinations with in-domain data for acoustic model training and out-of-domain language model training—and vice versa. However, the best word error rate of 38.5 % was achieved with in-domain training data only. The authors conclude that reasonable speech recognition accuracy for indexing and information retrieval can be achieved by exploiting out-of-domain training data in case of a lack of in-domain training data. However, even the word error rate achieved by the best system is still too high to replace human transcription.

Zajic et al. [2018] studied speech recognition with Czech MALACH interviews. The authors trained a hybrid DNN-HMM acoustic model using the Kaldi ASR toolkit with 84 hours of Czech MALACH recordings. A network with five hidden layers and sMBR sequence-discriminate training was applied. For evaluation, Zajic et al. used ten interviews with overall 60,000 running words. Using a language model that combines texts from MALACH with additional text resources, the authors achieved a word error rate of 42.0 % on their test data.

In an *oracle experiment*, Zajic et al. used the transcripts of the test data for language model training and achieved a word error rate of 19.5 %. The authors state that this oracle experiment shows the current performance upper bound of the speech recognition system for the author's data. On the one hand, the author's investigation demonstrates that even in such an oracle scenario and using back-then state-of-the-art speech recognition, systems still have a quite high word error rate on oral history data. On the other hand, in our view, the results of such an oracle experiment, in which test data are used for training a language model, should be interpreted with caution and should not be used for comparisons between systems.

Regarding the sources of errors and challenges of oral history interviews for speech recognition systems, Zajic et al. state, often in line with the aforementioned

works, that it is challenging to design an ASR system for oral history interviews that is accurate enough due to the nature of the interviews. The interviewees are usually elderly people, their spontaneous speech is frequently accented, and they are often emotional due to the nature of their experiences. Also, the speech quality is relatively low with many disfluencies and non-speech events such as crying and laughing. The regular use of colloquial words also negatively impacts speech recognition.

Towards the end of the presented research work, Picheny et al. [2019] proposed parts of data from the MALACH project as a new speech recognition challenge for English oral history. The authors are in line with the statement that the challenges of such interviews are still open problems for modern speech recognition systems. The reference results by Picheny et al. are produced by training hybrid DNN-HMM systems on 176 hours of manually annotated oral history interviews. The best system proposed is a hybrid LSTM-acoustic model with feature splicing and sequence-discriminative sMBR training achieving a 25.9 % word error rate without an LSTM-based language model and 21.7 % with an LSTM-LM.

At the beginning of this section, we described the limited interest of computer science in improving automatic speech transcription of oral history interviews. Moreover, comparing published works is difficult since most authors work with the data relevant to their individual use case. This is, at least in part, because no standardized, publicly available oral history data set with predefined training and test split was common—at least until the proposed data set of Picheny et al. [2019]. Unfortunately, however, even in 2021, the citation count of the proposed data set does not indicate an increase in research interest.

Another reason for the limited research interest could be the multitude of different challenges for ASR that often co-occur in oral history interviews. The challenges thus often cannot be clearly defined, making systematic analysis and improvement for an entire data set difficult. Many commonly used public ASR data sets and challenges focus only on specific challenges. Such notable ASR challenges until 2017 are CHiME-3 [Barker et al., 2015] and CHiME-4 [Vincent et al., 2017], which have investigated ASR in challenging noisy environments. However, this challenge comprises data that mainly was simulated and covered controlled recording conditions. Even for the *real* data portion of the challenge, read speech with the limited vocabulary of the WSJ corpus was used, which substantially limits the transfer to the oral history task.

In recent years, complex real-world challenges have been gaining interest, such as the *dinner party scenario* of the CHiME-5 challenge [Barker et al., 2018]. With the ever-increasing power of speech recognition systems, there is hope that someday these systems will also yield automatic transcription of oral history interviews with

sufficient quality. At present, however, improving the recognition quality of oral history interviews is still an open research field with numerous unsolved problems.

In conclusion, various research works have studied speech recognition for oral history interviews over the past two decades. Most authors agree that oral history interviews are still a major challenge for ASR systems and that the recognition performance of the systems is not sufficient to be used for subtitling or replacing human transcription. Although the authors studied different data, languages, and speech recognition approaches over the years, this is consensus. Due to poor word recognition accuracy, most applications are limited to knowledge retrieval and indexing. The reasons for this poor performance are quite diverse—but seem to be similar across different data sets and languages studied.

A statistical meta-analysis of these challenges presented in the related works would be a great asset to the research on speech recognition for oral history. However, the aforementioned works in this research area had to use a different data set for their experiments. Moreover, the transcription errors were mainly evaluated qualitatively, not quantitatively. Thus, the results are not statistically comparable for a meta-analysis with the available body of studies. Instead, in the following, we summarize and categorize the different challenges in a structured, non-statistical way and put them in the context of the German oral history data studied in the presented work.

3.3.2 Challenges of Oral History Interviews for ASR

Almost all research works referenced in Section 3.3.1 report diverse, often quite similar, challenges that oral history interviews pose for transcription systems. This is true even though these works studied different data sets in different languages. Thus, there appear to be all-encompassing, general challenges of oral history interviews for ASR that most data sets share. In the following, we aim to provide a structured overview of these challenges and put them in the context of the German data studied in the presented research work.

A brief overview of the challenges and differences between oral history and broadcast can be found in Table 3.1. We choose the categories *acoustics*, *speech*, *language style*, and *topics* since the influence of the challenges on the speech recognition in each category can be assigned almost exclusively to either the acoustic or the language model. In the further course of the work, this allows us to investigate the respective influence of the component for the interviews and select relevant aspects for improvement.

The major challenge we face with the German oral history use case is that we have almost no annotated and temporally aligned data available that we can directly use for training an acoustic model. Most of the referenced works use between 50 to 200 hours of annotated, in-domain oral history interviews to improve speech

Table 3.1: Overview of the differences between broadcast and oral history recordings that influence the accuracy of speech recognition systems. Acoustic conditions and speech in recordings are mainly modeled by the acoustic model of an ASR system, while the language model represents the language style and the topics.

	Acoustic Model		**Language Model**	
	Acoustics	**Speech**	**Language Style**	**Topics**
Oral History	wide range of recording qualities, often noise and room-reverberation	high speech rate variation, mispronunciations, accented speech or dialects	ungrammatical constructions, colloquial language, hesitations and repetitions	from everyday life to historical topics
Broadcast	professional, high-quality recordings	slowly speaking, clear pronunciation, Standard German	well-formulated phrases	mainly politics and news

recognition performance. For the German oral history use case, such amounts of data are currently not available. As described in Section A.2.3, the speech recognition models for the Fraunhofer IAIS Audio Mining system are specialized for and trained on broadcast data. To overcome the lack of suitable training data, we study the adaptation of data sets and models to improve speech recognition performance in the presented research. Therefore, we highlight the differences between this broadcast data used for training and oral history interviews. In the following, we further give a more detailed explanation for each of the four categories from Table 3.1.

Acoustics

Acoustics is a large scientific field with numerous subdisciplines. At its core, acoustics deals with the study of sound waves, their propagation, interference, and perception. As a subfield of acoustics, *electroacoustics* further deals with the recording of sound waves, cf. Kleiner [2013, pp. 5 ff.].

Several related works that studied speech recognition for oral history interviews report poor recording quality of oral history interviews as a challenge for speech

recognition systems. This applies both to early and current works. The researchers mention background noises as well as unsuitable recording conditions in this context.

For the German oral history data studied in the presented research work, the interviews were often recorded in the living rooms of the contemporary witnesses using common recording equipment. From a speech recognition point of view, this can be regarded as a problematic recording situation. Here we observe two problems regarding the recording quality:

1. Reviewing interviews from our German oral history collection leads us to the hypothesis that reverberation in the recording room is one of the main acoustic challenges for speech recognition systems in this domain. This seems to be due to an often large distance between the speaker and the recording microphone.

   Humans compensate room reverberation of small and medium-sized rooms quite well due to the *precedence effect* of binaural hearing. Sound waves reflected in the room arriving at the listener within 2 to 50 ms after the first wavefront are perceived by humans as a single auditory event and not as reverberation or echo, cf. Avan et al. [2015]—even if the subsequent, reflected wavefronts are louder. Asymmetric hearing loss is expected to impact this echo suppression negatively, and affected persons may experience difficulties in the presence of reverberation.

   Since our data is mainly single-channel (*mono*) recorded, we suspect that room reverberation has a great, if not greater, impact on recognition quality than the noise. However, reverberation is not explicitly identified as a challenge in the reviewed literature. Therefore, it is an open question whether reverberation is genuinely one of the main challenges of oral history interviews for speech recognition. We discuss and experimentally investigate this hypothesis for our interviews in Section 4.4.

2. The recording equipment changed during the years of recordings and resulted in a wide range of different sound qualities. As reported in Section 3.2.3, some recordings were conducted in the late 1970s. Some of these older interviews have a recording quality that can hardly be compared with today's devices. The wide range of different recording qualities probably also has a big impact on recognition accuracy. In a study in Section 5.5, we investigate the influence of this recording quality by comparing the primarily studied, older interviews with more recently recorded interviews with better recording qualities.

> Some oral history recordings in the archive have such a poor recording quality that the recorded speech is hardly understandable even for humans. However, we exclude these recordings from the investigations.

In the hybrid speech recognition approach, the acoustic model is the component that is susceptible to poor audio signal quality. As described in Section A.2.3, in the broadcast recordings used for acoustic model training, hardly any room reverberation is to be expected. Thus, we have a substantial mismatch between training data and application—and a lack of annotated speech for training from the target domain. In Chapter 4, we present our research work to minimize this mismatch by adapting the broadcast training data using noise and reverberation data augmentation. We give a detailed overview of the used training data and other studied corpora in Section 3.4.

Speech

In addition to acoustics, the speech of the interviewed persons also greatly influences the acoustic model of the speech recognition system. Often, the interviewed eyewitnesses have no professional background in giving speeches and spontaneously narrate their life stories. Therefore, spontaneous speech in oral history interviews is a challenge highlighted—directly or indirectly—in various aforementioned works.

Spontaneous speech poses various challenges for speech recognition systems, cf. Ward [1989]; Dufour et al. [2014]. In particular, spontaneous speech may result in pauses filled with speaker noises and mispronounced words. It is common to assume that if the person is speaking very fast, wrong or indistinct pronunciations may occur more frequently and pose a challenge for speech recognition systems. Among other acoustic features, Dufour et al. [2014] studied the speech rate and the speech rate variations (in terms of the standard deviation of phone duration) as an attribute of spontaneous speech. The authors found a correlation between spontaneity and both speech rate attributes. We study the speech tempo for the German oral history interviews in comparison to German speech from other domains in Sections 3.4.11 and 3.4.12.

Depending on the origin of the interviewees, dialects or accented speech can be an additional challenge for an acoustic model—and the pronunciation lexicon as a sub-component—trained on the standard variation of the language only. Since eyewitnesses are often elderly persons, some works mention the age of the interviewees as a potential challenge for speech recognition systems. Here both age- and health-related changes in the way of speaking could lead to the indistinct pronunciation of words that pose additional challenges for the acoustic model. These challenges depend on the interviewees and are not expected to appear to the same extent in every data set.

In contrast, broadcast recordings usually contain planned (or prepared) speech. The persons are often professional speakers who focus on clear and well understandable pronunciation—since the broadcasters intend the speech to be well understood by the audience. In Chapter 5, we aim at overcoming this mismatch of broadcast training data and oral history interviews by applying a transfer learning approach in acoustic model training. In Chapter 6, we further propose and study cross-lingual acoustic model adaptation to further overcome the lack of training data in the target domain and language.

Language Style

Spontaneous speech does not only have a negative influence on the speech but also negatively influences the language style or *register*, cf. Ward [1989]; Dufour et al. [2014]. It can lead to ungrammatical sentence constructions, such as word or partial word repetitions, sentence restarts, or interjections. If the language model is not designed to handle such challenges, it may degrade recognition performance.

Furthermore, colloquial language and colloquial words the interviewees use can negatively impact speech recognition if the language model can only model them inadequately. In the case of Audio Mining, where the language model is trained on news text and articles—thus, mainly containing well-formulated phrases—we expected this to be a major challenge. Therefore, we use intrinsic evaluation metrics to study the influence of broadcast language models on oral history speech recognition in Section 3.5.

Topics

A wide range of topics is expected in oral history interviews since interviewees are often encouraged to narrate their entire life story or various life experiences. Depending on the archive in question, there may well be a focus on topics—such as the Second World War or the postwar period. However, even within this topic focus, a wide variety of subjects is expected due to different life experiences individuals face in everyday life. This wide range of topics can pose a challenge for language models.

For broadcast recordings, there is also a wide range of different topics. However, a large focus on politics and the daily news is to be expected. Thus, the frequently used topics and phrases differ greatly from eyewitnesses' life accounts in oral history interviews.

3.4 Corpora for Automatic Speech Recognition

Oral history is the primary use case to be examined in this research. Therefore, a representative data set is necessary for evaluation, which will be presented in the further course of this section. In ASR research, it is common to focus on optimizing the error rate on a single data set, beating the previous benchmark. However, the proposed models in the presented research work will be applied in productive, real-world Audio Mining systems, and high robustness of the models must be assured. The call for robustness in this work is reinforced by the fact that oral history interviews have a higher range of different challenges than many other common speech recognition applications—as discussed in Section 3.3.2. Therefore, we consider an evaluation solely on a data set from the oral history domain insufficient to make a feasible statement about its applicability on real-world data. To overcome these limitations and ensure the models' high robustness even in situations for which the model has not been trained, we will report results on multiple different test sets from diverse domains. These sets serve as a control group for the trained model. Through this approach, we expect to better estimate the performance of the proposed models for unseen, real-world applications in oral history archives—and maybe beyond.

Evaluating a speech recognition system on multiple data sets from different domains is not common in ASR research—but is consistent with some recent work in the literature. For instance, Likhomanenko et al. [2021] advocate rethinking evaluation in ASR and examining systems on several evaluation sets from different benchmarks to estimate performance for real-world data and detect systems that suffer from *domain overfitting*. Szymański et al. [2020] also raise related criticism. The researchers express their skepticism towards very low word error rates on single, well-known ASR benchmarks published in recent research. These results fuel overly optimistic expectations of speech recognition systems that are not fulfilled for real-life situations. Szymański et al. argue that, contrary to popular belief, current speech recognition systems cannot satisfactorily transcribe spontaneous human conversations.

In the following, we present the different ASR corpora we use in the presented research work. Some of the data sets were available early on, while other data sets were developed or made available in the course of the years. As a contribution, we present the German oral history evaluation set developed as part of the presented research work and proposed in [Gref et al., 2018a] to investigate this use case. Additionally, we analyze different properties of the several data sets in more detail to better contextualize the oral history use case studies and the discussed challenges of the previous section. We end the section reporting state-of-the-art results from the literature on the data sets that were recent at the beginning of the presented research work.

Table 3.2: Overview of the several ASR data sets used throughout the presented research work with fundamental information. Data sets from the broadcast domain are presented in the two top parts: a training set and different evaluation sets, respectively. In the lower part, individual test sets from other domains are presented.

Data Set	Length [hrs:min]	Num. of Segments	Num. of Words	Unique Words
GerTV1000h (Training Set)	991:53	773,631	9,406,119	243,313
DiSCo Planned Clean	0:55	1,364	9,184	2,939
DiSCo Spontaneous Clean	1:55	2,861	20,740	4,019
DiSCo Planned Mix	1:27	2,200	13,698	4,083
DiSCo Spontaneous Mix	1:06	1,650	12,071	2,764
German Broadcast 2016	1:01	227	10,143	2,796
Challenging Broadcast	1:45	593	17,354	4,179
Oral History	3:31	2,392	27,708	4,582
Interaction (Linguistics)	0:48	2,630	10,015	1,392
Spoken QALD-7	0:15	212	1,467	624

3.4.1 Overview

An overview of the different corpora is given in Table 3.2. The top two parts of the table present data sets from the broadcast domain—a large-scale broadcast training set and several broadcast evaluation sets, respectively. Since broadcast is the core application domain of the IAIS Audio Mining system, various German in-house broadcast data sets with different attributes were introduced throughout the years. The lower part of the table presents three individual test sets from other domains: oral history, interaction, and question-answering systems. Some data sets were annotated in-house with the *ELAN annotation tool for audio and video recordings*[4] [Brugman and Russel, 2004]. Other data sets were annotated by clients or project partners themselves or commercial annotation vendors. In the following, we give a brief description of each data set in the order listed in Table 3.2.

[4]Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, https://archive.mpi.nl/tla/elan

3.4.2 The GerTV1000h Training Set

Since 2014, the Fraunhofer IAIS has used the in-house, 1000 hour large-scale German broadcast corpus *GerTV1000h* [Stadtschnitzer et al., 2014] to train acoustic models for the Audio Mining system. We observe state-of-the-art speech recognition performance of such models for broadcast data on a regular basis, cf. Stadtschnitzer [2018]. At the beginning of the present research, GerTV1000h was probably one of the largest transcribed German data sets for automatic speech recognition.

Since the recordings are from the broadcast domain, they are usually recorded and post-processed by professional sound engineers using professional equipment. The recordings generally have no or only slight background noise, barely perceptible reverberation, and the volumes are well adjusted.

As an ASR training data set, GerTV1000h is annotated somewhat differently than the evaluation corpora. In addition to the verbatim transcription, it also contains distinct symbols for speaker noises, hesitations, and unintelligible words, cf. Stadtschnitzer [2018, p. 45–46]. These symbols are mapped to distinct phone models during acoustic model training. This aims to separate such sounds and noises from the speech—and ultimately improve word recognition.

3.4.3 Difficult Speech Corpus (DiSCo)

The Difficult Speech Corpus (*DiSCo*) [Baum et al., 2010] is an evaluation corpus for the German broadcast domain. It has been split into four evaluation subsets: planned and spontaneous speech, each in clean and mixed acoustic conditions. Each of these evaluation subsets contains roughly between 1 and 2 hours of annotated speech. Although the subsets are labeled *spontaneous* or have *mixed* acoustic conditions, the distortions are generally weaker in the broadcast domain than in oral history or other domains.

The DiSCo subsets, in particular the clean subsets, were the main evaluation sets used by Stadtschnitzer [2018] for the long-term development of the German Broadcast Audio Mining system between 2012 and 2018. Due to this fact, this evaluation set is one of the first evaluation sets available for the presented research work and was used in all experiments from the early beginning. Initially, results were often reported and published separately for all four subsets. In the course of the work, when more and more representative and meaningful data sets became available, the significance of the results on DiSCo for the research questions decreased. Thus, in later publications, only the mean value of the four subsets was often reported.

3.4.4 German Broadcast 2016

German Broadcast 2016 is another in-house evaluation set for the German broadcast domain. As the name implies, it was introduced in 2016. The data set can be thought of as a blend of the four scenarios from the DiSCo evaluation subsets as it is usually found in actual broadcasting programs. The 20 different recordings in the data set include both studio recordings of newscasters in excellent recording quality and without background noise, as well as outdoor interviews with passers-by in the presence of street noises. There are also occasional voice-overs of German speakers over foreign speakers' voices—a particular challenge for speech recognition systems since two different voices are simultaneously present. German Broadcast 2016 has about the same size as DiSCo planned clean but substantially fewer segments. This indicates that the segments are usually much longer, as discussed in the analysis in Section 3.4.11.

3.4.5 Challenging Broadcast Evaluation

The *Challenging Broadcast* evaluation set was available in late 2018 and is similar to the most challenging DiSCo subset: Spontaneous Mix. The data comprises 1.75 hours of German broadcasts in total with 16 different recordings—eight from radio and eight from TV broadcasts. The composition of the data sets is quite diverse. It contains quite conventional broadcast recordings with planned speech, similar to DiSCo or German Broadcast 2016. However, it also includes several highly challenging recordings characterized by a lot of spontaneous speech in challenging acoustic conditions, including overlapping and sometimes even dialectal speech.

3.4.6 Proposed German Oral History ASR Test Set

In [Gref et al., 2018a], we proposed a *German Oral History* ASR test set to measure the performance of the developed automatic speech recognition systems for oral history interviews. This test set is a subset of the *Deutsches Gedächtnis* archive (cf. Section 3.3.1), representing the archive's 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 includes early interviews as well as recently conducted ones and represents the interview methods of different academic disciplines. The recordings used for the test set took place between 1980 and 2012. With respect to gender and age, the selection aims at representing the entire collection.

Within these criteria, the test set was randomly selected by the project partners from the *Deutsches Gedächtnis* archive in the KA^3 project. Overall, the test set

contains interviews from 35 different speakers with an overall length of about 3.5 hours, with more than 27,000 spoken words and a vocabulary larger than 4,500. This makes the oral history test set one of the most comprehensive ASR test sets at Fraunhofer IAIS.

The values described here and in the further course refer to the latest version of the oral history test set used unchanged for large parts of the presented research work. In some of our initial experiments, we used a *non-segmented* version of the test set with an above-average segment length of a few minutes. We found in later investigations that this segment length is challenging for LF-MMI models and that the performance obtained with this segmentation is not representative of our real-world applications in Audio Mining. The experiment on the influence of the segment length is presented in Section 4.3.2 in the next chapter.

3.4.7 Interaction (Linguistics)

The *Interaction* test set represents the linguistic use case of the KA^3 project where human interaction between persons is of interest. Thus, the Interaction set contains recordings of people informally talking to each other about different topics. Generally, this set is characterized by very fast, partly overlapping, highly colloquial spontaneous speech, speaker noises such as laughter, and unclear pronunciation. From the perspective of ASR, the recordings took place in challenging acoustic conditions with a far distance of the speakers from the microphone. The Interaction test set was completed in late 2018 and used for further research from this time on. Despite its comparatively small size with only 48 minutes, it is striking that the set contains a similar number of segments as the much larger oral history test set (3:31 hours) or DiSCo Spontaneous Clean (1:55 hours). This indicates that the average segment length is probably much smaller than the other data set.

Furthermore, the Interaction test set has more running words than data sets of similar sizes, such as DiSCo Planned Clean. Although this is not an exact metric, as will be shown later, this observation indicates the subjective characterization of the above-average speed of speech of the persons in the recordings. The multitude of these challenges makes this corpus probably one of our most challenging test sets. The general properties of the data set are further explored in Section 3.4.11.

3.4.8 Spoken QALD-7

The *Spoken QALD-7* corpus contains in-house recorded questions for a question answering system based on prompts from [Usbeck et al., 2017]. Each prompt is a question that targets the property of one particular entity—a well-known person or object. Entities tend to be poorly modeled in most language models for continuous speech recognition since they rarely occur in continuous texts compared

to function and everyday words. Therefore, the main challenge of this test set for speech recognition is posed by the presence of the entities in each prompt. This also leads to a comparatively large vocabulary compared to the running words observed in Table 3.2.

Several speakers recorded the data in 2018 using a web interface and their respective recording setup—a headset or built-in laptop microphone, for instance. Thus, the recording quality in the data set is quite heterogeneous. One segment usually comprises exactly one question prompt. Overall, 210 different prompts are present in the 212 recordings.

With a total length of only 15 minutes and just under 4100 running words, Spoken QALD-7 is the smallest test set in the collection. Since it only contains read speech, it is only of minor significance for investigations regarding spontaneous speech. However, it is a valuable addition to the analysis due to the heterogeneous recording conditions, the different domains, and the rich vocabulary in relation to the number of running words.

3.4.9 Raw, Transcribed Oral History Interviews for Forced Alignment Experiments

In Chapter 5, we use automatic transcript alignment of transcribed but not time-aligned oral history interviews to overcome the lack of German oral history training data. Randomly selected oral history interviews of contemporary witnesses provided by the archive *Deutsches Gedächtnis* serve as data for these experiments. Since the recordings are not segmented nor time-aligned and often not transcribed verbatim, they cannot be directly used for speech recognition—and thus are not listed in Table 3.2.

The primary criterion in selecting the interviews was that the interviewees did not appear in the oral history test data. The interviews were conducted in 1982–2015 using varying recording equipment in different recording conditions. The recording quality is quite similar to the oral history test set. However, it was not ensured in advance that the recordings are understandable to humans—which it was for the test set. Thus, the recording quality differs from barely intelligible to good quality with only slight distortions. The latter are primarily low-energy noises from cassette recording machines or light reverberation due to a large distance to the recording microphone in a medium-sized room. Some speakers have a dialect or accent, but most speak High German.

Different historians throughout many years transcribed these interviews using varying transcription styles and formats. It cannot be guaranteed, nor was it assured in advance that the entire transcription of the interview is present for alignment or whether it is correct. The final data set comprises 150 interviews.

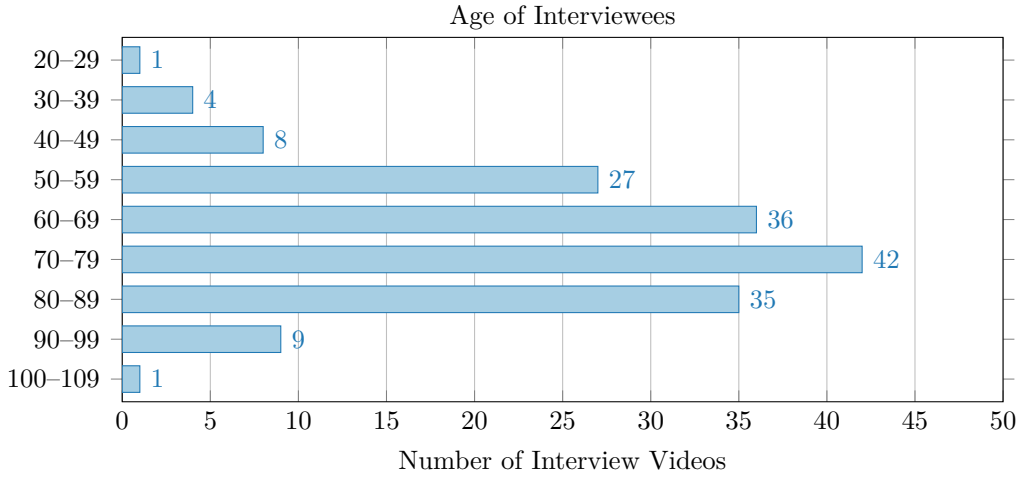


Figure 3.1: Age distribution in the HdG interview data set. No age information was available for one of the 164 videos.

These were not all provided and processed at once but gradually in four tranches in 2018–2019.

3.4.10 The HdG Oral History Data Set

An additional oral history data set became available in the final phase of the present research in 2021. This data set was created as part of the research project *Multi-Modal Mining of German Oral History Interviews for the Indexing of Audiovisual Cultural Heritage*, in cooperation with the *Haus der Geschichte* (HdG) foundation. The data set is a representative collection of the HdG *Zeitzeugenportal*, cf. Section 3.2.2. With the corpus, the project aims not only to investigate automatic speech recognition. In particular, it is used to study and develop more complex search and indexing technologies beyond ASR, such as multi-modal recognition of perceived emotions and sentiment [Gref et al., 2022a].

We selected 10 hours of German oral history interviews from the HdG *Zeitzeugenportal* for our experiments. The *HdG data set* comprises 164 different interview videos of 147 distinct interviewees. On average, the interviews in our data set have a length of 3.6 minutes. The selected interviews were recorded between 2010 and 2020. Thus, the selection is representative of the more recent videos on the portal. It includes 66 interviews with professional speakers, who pursue a representative profession, and 98 interviews with non-professional speakers. The data set is not published and is only used in-house due to the General Data Protection Regulation and the personal rights of the interviewees.

In addition, we aimed to represent different emotions in the selection of videos and create a heterogeneous data set in terms of age and gender. The age distribution of interviewees in the data set is shown in Figure 3.1. We have a strong focus on interviewees between the age of 50 to 89 years as these are the most frequent interviews in the archive. Nevertheless, we have deliberately included videos of younger and older interviewees. Throughout the entire archive, male interviewees make up most of the data. However, a representative selection would substantially underrepresent female interviewees. Therefore, we have included additional female speakers in the HdG data set. Overall, the HdG data set contains videos with 104 male and 60 female interviewees.

The HdG portal's interviews are somewhat different from the interviews in the (KA^3) Oral History test set we utilize in the entire presented research work as the primary object of study, cf. Section 3.4.6. While the archive *Deutsches Gedächtnis*, who provided the KA^3 interviews, works with full-length interviews with an average length of 3.5 hours, the HdG *Zeitzeugenportal* shows thematic clips of interviews with a length of 3 to 5 minutes. The interviews in the *KA^3 data set* were recorded between 1980 and 2012 and represent the archive's wide range of interviews. Since the interviews in our HdG data set are recorded in more recent times, often with more recent or professional equipment, the HdG data set overall have better audio recording quality than the KA^3 data. In other aspects, such as language style, age of the interviewees, dialects, and topics, the HdG and KA^3 data are quite similar.

We apply the HdG data set in two experiments in 2021–2022. In Section 3.6, the data set is the primary study object used to estimate a human word error rate for the transcription of oral history interviews. In this experiment, the reference transcription for later ASR evaluation is generated by three different persons. These transcripts are not merged, resulting in three different transcriptions for the entire data set. More information on the data set transcription is presented as the experiment results in the respective section.

After the annotation and transcription, the HdG data set is split into speaker-independent training, development, and test subset for model training and evaluation. An overview of these sets is given in Table 3.3. Overall, 358 segments with roughly 0.5 hours could not be annotated and were removed from the data set. These were, for example, segments containing intros, fade-ins, or pauses.

The transcribed data set is then used in Section 5.5 for comparative acoustic model adaptation experiments to assess the influence of the acoustic recording conditions in the oral history domain. Furthermore, the evaluation and comparison of the two different oral history data sets help to obtain a reliable estimate of the recognition performance of our proposed models for real-world applications. The difference in performance between the two data sets may provide insights into

Table 3.3: Overview of HdG oral history data sets after annotation and split into speaker-independent subsets. The primary (KA^3) Oral History test data set from Section 3.4.6 is included for comparison. The length is reported as [hrs:min].

Set	Videos	Segments	Length
HdG Training	104	1,863	6:21 h
HdG Development	27	430	1:26 h
HdG Test	33	471	1:44 h
KA^3 Test	35	2,392	3:31 h

the impact of audio quality and video age of oral history interviews for speech recognition.

3.4.11 Statistical Analysis of the ASR Data Sets

The previous section has already foreshadowed different statistical properties of the described data sets, such as the average segment length, that can be inferred from the raw data. These properties can impact speech recognition performance, and a comparison of the properties between the data sets can help to understand these effects better. Thus, for a more precise interpretation of the data, three values have been derived for each data set from the raw data and are summarized in Table 3.4: the average segment length, the average number of running words per segment, and the words per second—as a rough, first approximation of the speed of speech or speech rate.

The segment length varies greatly for the different data sets. The only exceptions are the DiSCo subsets, which have relatively similarly distributed segment lengths. As suspected in advance, the Interaction test set has the shortest segments with just 1.1 seconds on average. German Broadcast 2016 has the longest segments with an average of just under 16 seconds. Most data sets have a standard deviation below 3.5 seconds in segment length, indicating a fairly homogeneous segmentation. However, exceptions to this are German Broadcast 2016 and Challenging Broadcast, which have a fairly high standard deviation, thus, inhomogeneous segmentation.

In general, segments which are too long can pose problems for HMM-based speech recognition systems since the number of active states during decoding is limited. Very long segments increase the size of the decoding lattice, which can complicate decoding. However, up to a segment length of approximately 30 seconds or slightly higher, we have not observed a noteworthy negative impact on the recognition performance. The distribution of segment lengths indicates that a

Table 3.4: Statistics properties for the several ASR data sets used throughout the presented research work. Reported values are the arithmetic mean ± the standard deviation calculated across all segments of the data set. The segment length is reported in seconds.

Data Set	Segment Length [s]	Words Per Segment	Words Per Second
GerTV1000h (Training Set)	4.6 ± 2.0	12.2 ± 6.1	2.7 ± 0.8
DiSCo Planned Clean	2.4 ± 1.5	6.7 ± 4.9	2.7 ± 0.8
DiSCo Planned Mix	2.4 ± 1.3	6.2 ± 4.0	2.6 ± 0.8
DiSCo Spontaneous Clean	2.4 ± 1.6	7.2 ± 5.5	2.9 ± 0.9
DiSCo Spontaneous Mix	2.4 ± 1.4	7.3 ± 5.4	2.9 ± 1.1
German Broadcast 2016	16.1 ± 12.6	44.7 ± 34.7	2.9 ± 0.6
Challenging Broadcast	10.6 ± 17.0	29.3 ± 44.9	3.0 ± 1.2
Oral History	5.3 ± 3.5	11.6 ± 10.0	2.1 ± 0.8
Interaction (Linguistics)	1.1 ± 0.8	3.8 ± 2.9	3.6 ± 1.4
Spoken QALD-7	4.3 ± 1.3	6.9 ± 2.1	1.7 ± 0.4

large proportion of segments are substantially below this value. Only for segment lengths of a few minutes we observe that the recognition can be negatively affected. This can vary depending on the type of acoustic model, as we have investigated in Section 4.3.2 for LF-MMI and cross-entropy trained models.

Segments that are too short are not a problem per se. However, it is evident that a strong correlation between the segment length and the number of words per segment exists for speech segments. Thus, short segments usually only contain a few words. Since the speech recognition system processes segment by segment and does not take into account the context of neighboring segments, very short segments can pose challenges to the language model appropriately estimating word sequence probabilities—and ultimately decrease speech recognition accuracy. According to Table 3.4, the DiSCo subsets have the shortest segments and the least number of words among the broadcast sets, with an average of 6–7 words and a standard deviation of 4–5 words per segment. However, the Interaction set has just under four words and a standard deviation of just under three words. For such short segments in the Interaction test set, a 5-gram language model, therefore, has hardly any chance of unfolding the full potential of its modeling capability.

In Section 3.3.2, we elaborated on the challenges of oral history interviews for speech recognition systems and that spontaneous speech is one of the suspected challenges. Dufour et al. [2014]'s work has shown that spontaneity of speech correlates with both the speech rate and the variation in speech rate. The words per

second are used in Table 3.4 as a first rough approximation of the speech rate to evaluate spontaneity for our corpora. As expected, we observe a slight increase in the average words per second as well as the standard deviation for the Spontaneous DiSCo subsets compared to the planned ones. Another increase is observed for Challenging Broadcast, where spontaneous speech recordings are more prevalent. The Interaction test set has the highest average words per second and the highest standard deviation, which indicates a very high rate of speech and variation.

It is somewhat surprising that the words per second for the oral history test set are lower than for the broadcast test sets. The standard deviation is also within the range of planned broadcast speech. This leads us to the hypothesis that a high speech tempo or speech tempo variance common for spontaneous speech is not predominant for our interview collection—contrary to what is implied in the related works for other oral history corpora. However, a definitive conclusion cannot be given at this point since the approximation of the speech rate via counting words per time unit is subject to some apparent limitations. To overcome these limitations and estimate the speed of speech more precisely, in the following, we investigate the estimation of the speech rate via the phone rate for our data sets.

3.4.12 Phone Rate Estimation of the ASR Data Sets

Estimating the speech tempo with *words per second* is easy to implement and requires little computational effort since it is just a simple statistical evaluation of the given ASR annotation. However, the approach is highly dependent on segmentation and transcription. The estimated speech rate will erroneously decrease if more speech pauses are included at the segment's beginning or end. Longer pauses between words also decrease the estimated speech rate since the overall segment length is increased. Ideally, these pauses would need to be removed by alternative segmentation. Particularly in German, word compounds are an issue as well. Words that consist of several compound nouns can be very long—but are weighted the same as very short words in the *words-per-second* approach.

To better estimate the speech rate and rule out these issues, we use a phone rate estimation based on an alignment of the reference transcription with GMM-HMM models. This approach is more complex because trained acoustic models are required. We use GMM-HMM LDA+MLLT+fMLLR acoustic models trained as part of the Kaldi DNN-HMM bootstrap training in the later Chapter 4. Usually, these models are used to align the reference transcript of the training data for the neural network training in the last step of the bootstrap. We apply these models on the test sets to obtain the duration for each spoken phone, cf. [Dufour et al., 2014]. For a given segment, the phone rate is calculated by dividing the number of non-silent phones spoken in the segment by the overall total spoken duration of

these non-silent phones. With N as the number of non-silent phones, this is

$$\text{Phone Rate} := \frac{N}{\sum_{n=1}^{N} \text{Duration of non-silent phone } n},$$

which is equal to the reciprocate of the arithmetic mean of the duration of all non-silent phones in the segment (average phone duration). As discussed above, we consider only non-silent phones to avoid the phone rate being subject to the chosen segmentation, making the speech rate of the differently segmented data sets comparable.

The estimated phone rates for all German data sets used in the presented research work are presented in Table 3.5 in comparison to the words per second from the last section. We report the results using three different GMM-HMM systems from the later Chapter, cf. Section 4.4.5. Overall, the phone rate and the words-per-second approaches for estimating the speech tempo show similar trends. This indicates that the words-per-second approach is quite sufficient for preliminarily and qualitatively estimating the average differences of speech tempos between our test sets with simple statistical calculation. However, it is also evident that the different annotations of the test sets certainly influence the relationship between phone rate and words per segment.

The phone rates are different for each model but consistent for the individual test sets. The spontaneous broadcast test sets have a higher phone rate and standard deviation than the planned ones. Challenging Broadcast still is the test set with the highest mean speaking rate and speaking rate variance among the broadcast test sets. The highest overall speaking rate and variance has Interaction—as previously suspected based on the words per second. The Oral History and Spoken QALD-7 sets have an overall low speaking rate—both in terms of the words per second and the phone rate. The phone rate is lower than even the planned broadcast sets. This indicates a fairly leisurely speaking rate, as estimated with the words-per-second approach.

The differences in the phone rates of the three models have two primary reasons that are both rooted in how the models are trained. First and foremost, the three models show slightly different behavior when determining the start and end times of phones. A sample-based, subjective evaluation revealed that the *Clean* model is more precise in separating non-silent from silent phones, such as speech pauses, speaker noises, and fillers. Thus, the duration of non-silent phones by the Clean model is shorter, resulting in a higher phone rate. The two 3-fold models, as described later in Section 4.4, were trained using speech with additive noises and room reverberation. As discussed later in Section 4.4.1, room reverberation leads to a *smearing* of speech along time and a slower decay of signals, making

Table 3.5: Phone rate estimation using different GMM-HMM LDA+MLLT+fMLLR models for phone alignment. The phone rate is reported as the arithmetic mean ± standard deviation per data set in *phones per second*. Segments or words that the respective model could not align, for instance, due to wrong labeling or pronunciations, were omitted for the estimation. The words per second from Table 3.4 are reported for comparison.

		Phone Rates		
Data Set	**Words Per Second**	**Clean**	**3-fold v1**	**3-fold v2**
GerTV1000h	2.7 ± 0.8	14.5 ± 2.4	14.5 ± 2.4	14.1 ± 2.5
DiSCo Planned Clean	2.7 ± 0.8	13.8 ± 2.7	13.8 ± 2.7	13.3 ± 2.7
DiSCo Planned Mix	2.6 ± 0.8	13.8 ± 2.5	13.7 ± 2.5	13.3 ± 2.5
DiSCo Spontaneous Clean	2.9 ± 0.9	14.4 ± 3.4	14.4 ± 3.4	13.9 ± 3.3
DiSCo Spontaneous Mix	2.9 ± 1.1	15.0 ± 4.0	14.7 ± 3.9	14.6 ± 3.9
German Broadcast 2016	2.9 ± 0.6	14.9 ± 2.0	14.9 ± 2.0	14.6 ± 2.1
Challenging Broadcast	3.0 ± 1.2	15.4 ± 4.1	15.4 ± 4.2	14.9 ± 3.9
Oral History	2.1 ± 0.8	13.1 ± 3.3	12.8 ± 3.2	12.7 ± 3.1
Interaction (Linguistics)	3.6 ± 1.4	17.6 ± 6.6	15.7 ± 5.7	17.1 ± 6.3
Spoken QALD-7	1.7 ± 0.4	13.2 ± 2.1	13.2 ± 2.0	12.7 ± 2.1

differentiation of speech and non-speech parts more difficult and also ambiguous. Since the 3-fold models have been bootstrap-trained primary with such reverberated data, it is evident why these models are considering more parts of the signal to be speech. This results in greater phone durations and, thus, ultimately lower phone rates.

The second but less substantial reason for the difference in the phone rates is that the different models cannot align all segments and omit certain segments, which are ignored for the average phone rate calculation of each test set. The reasons for the failed alignments are manifold, such as incorrect reference transcription, wrong entries in the pronunciation lexicon for certain words, and challenging acoustics conditions. Figure 3.2 presents the share of segments that the different models could not align. Obviously, for the challenging data sets, more segments cannot be aligned. The 3-fold models are more robust than the Clean model and can align more segments, particularly of the data sets with challenging acoustics, such as Challenging Broadcast, Oral History, and Interaction. The absolute numbers of aligned segments are documented in Table B.1 in the Appendix.

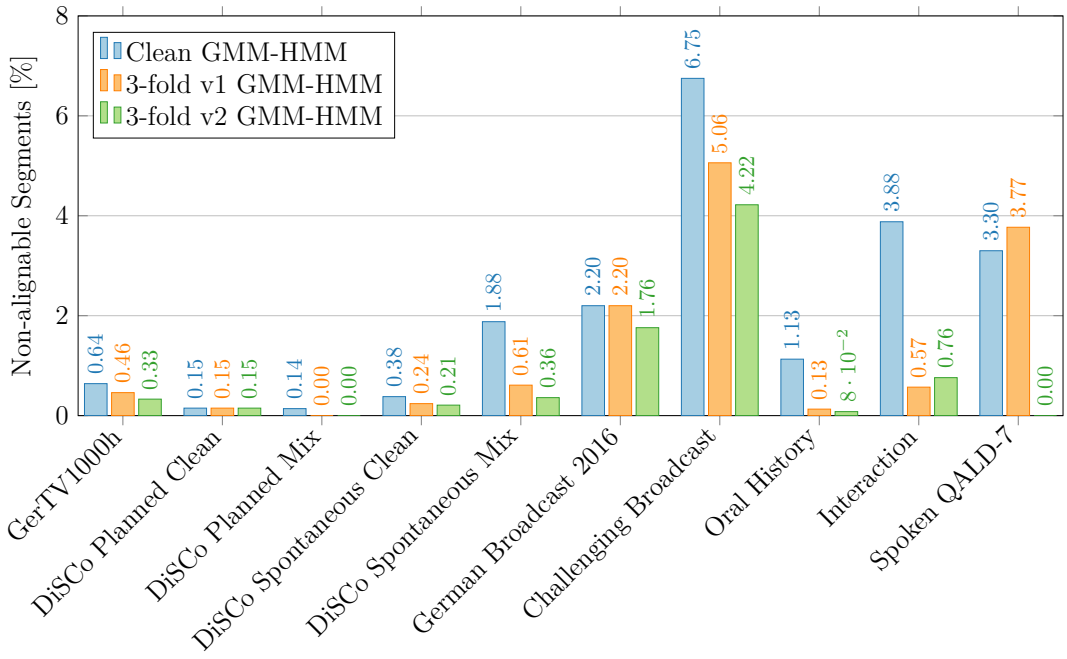


Figure 3.2: Percentage of segments that could not be aligned by the HMM acoustic model for phone rate estimation. The 3-fold GMM-HMM models are trained more robustly than the Clean model, cf. Section 4.4. They thus can align more data, particularly in challenging acoustic conditions.

Overall, the phone-rate experiments confirm the hypothesis stated in the previous section based on the words per second. A high speech tempo or speech tempo variance that is usually common for spontaneous speech is not predominant for our oral history interview. The values are roughly in the range of planned broadcast speech. Thus, special consideration of the speech rate for our experiments, e.g., with customized speech perturbation techniques, is therefore not necessary. The other discussed challenges, such as acoustic recording conditions, require greater consideration. However, our results do not imply that spontaneous speech is not a challenge in oral history interviews per se. Moreover, spontaneous speech manifests itself differently in the interviews, e.g., through longer speech pauses, hesitations, rephrasing, and ungrammatical sentence constructions.

3.4.13 Baseline Model and State-of-the-Art Results at the Beginning of the Presented Research Work

The Fraunhofer IAIS Audio Mining System has been further developed over the years by several researchers investigating and improving different aspects of speech recognition or speech signal analysis. Before and during the beginning of this re-

search work, this was mainly the aforementioned work by Stadtschnitzer [2018], who studied *robust speech recognition for German and dialectal broadcast programs*. Furthermore, at that time, Fraunhofer researchers studied end-to-end speech recognition for German broadcast using CTC-RNNs (cf. Section 2.3) with the Eesen toolkit [Miao et al., 2015a]. The models available at the beginning of the presented research work serve as a baseline for the proposed approaches.

Evaluation Measure: The Word Error Rate and Its Limitations

In speech recognition, ASR systems' performance is usually measured in terms of the *word error rate* (*WER*). The word error rate is derived from the Levenshtein distance (or edit distance), cf. Levenshtein [1965], using words instead of characters. In particular, for ASR, the edit distance is the minimum number of word edit operations (word insertions I, word deletions D, and word substitutions S) required to transform the reference word sequence to the hypothesis—the word sequence generated by the ASR system in question. The word error rate is the quotient of the edit distance and the total number of words in the reference N:

$$\text{WER} = \frac{I + D + S}{N}. \tag{3.1}$$

Obviously, the edit distance and, ultimately, the word error rate are highly dependent on the reference transcription. Since this reference is usually created by humans, it may be subject to some limitations. It is not only necessary to take into account that humans can make mistakes when transcribing. Unlike some annotations in other machine learning domains, the transcription of spontaneous speech has a high degree of inherent ambiguity, cf. Stolcke and Droppo [2017]. In our work, we observe that different (correct) spellings of words in German can lead to ambiguity, such as noun compounds that are sometimes written as compounded, hyphenated, or separated words. Also, in spontaneous speech, humans often tend to transcribe what they understand and not necessarily what was actually said. It seems that speech errors are often unconsciously overheard and corrected. Speech recognition systems are usually more precise in this respect and transcribe mispronounced words, repeated words, and slips of the tongue precisely as they were uttered.

It depends on the further usage of the transcript, whether such a verbatim, phonetically exact transcription or more human-like, corrected transcription is to be considered *correct*. This also applies, for example, to the transcription of hesitations. For linguistics and specific historical research questions, a verbatim transcription as exact as possible of what was actually said is relevant. A human-like, corrected transcription is more desirable for other applications, e.g., for subtitling videos or further processing by NLP systems. We are currently striving for pho-

netically exact transcription for our current task, including hesitations, as such transcriptions help researchers to investigate not only *what* was said but *how* it was said. Correction of the transcript is performed in subsequent post-processing steps, which we do not consider part of the ASR.

These issues are reflected, among others, when comparing the differences between human transcribers on the same speech data. Human transcription was studied, for instance, by Xiong et al. [2017a] and Saon et al. [2017] for the English Switchboard corpus [Godfrey et al., 1992] to determine a *human word error rate* and also to uncover correlations between transcription errors by humans and ASR systems on this data. Inspired by these works, we conduct experiments in Section 3.6 to estimate the human word error rate on oral interviews to address this issue for the oral history transcription task.

Another issue with the word error rate is that it only considers the total number of errors—but does not weigh different errors. A substitution is counted equally, regardless of a word was recognized entirely wrong, or only one letter in the recognized word is different. For our evaluation, we consider upper and lower case spelling. Thus, casing errors by the ASR are counted as common substitutions in the word error rate. However, since the word error rate is by far the most common metric in ASR research, and no other metric has yet become widely accepted for evaluation, we also use it to evaluate our experiments.

Baseline Results

As described in Section 3.4, the primary evaluation sets available at the beginning of the presented research work were the two clean subsets of the DiSCo corpus: *DiSCo planned clean* and *DiSCo spontaneous clean*. Furthermore, in 2016 the *German Broadcast 2016* evaluation set was introduced. The word error rates on these test sets of different acoustic models proposed until the start of this research are summarized in Table 3.6. The results of the DNN-HMM and pDNN-HMM (p-norm DNN-HMM) models have been published by Stadtschnitzer [2018, p. 64]. The results on CTC-RNNs experiments have been published by Schmidt et al. [2016]. All these models were trained on the entire GerTV1000h corpus. For evaluation, a *default* language model was used that is presented in detail in Section 3.5.

The DNN-HMM system from Table 3.6 was used as the recent acoustic model in the Fraunhofer IAIS Audio Mining system at the beginning of the presented research work. It achieves a 15.5 % word error rate on clean planned broadcast speech and 19.7 % on clean spontaneous broadcast speech. Using the more advanced acoustic model approaches improves recognition performance on these data by up to 3–4 percentage points. A further improvement is achieved by decreasing the pruning threshold of the speech model.

Table 3.6: Word error rates of acoustic modeling approaches at the Fraunhofer IAIS published in the literature. All models are trained on the entire GerTV1000h corpus. The DNN-HMM model marked with * is the acoustic model that was used in the Audio Mining system at the beginning of the presented research work.

Acoustic Model	Pruning Threshold	DiSCo Clean Plan.	DiSCo Clean Spont.	German Broadcast 2016
DNN-HMM*	1e-7	15.5 %	19.7 %	19.8 %
pDNN-HMM	1e-7	13.3 %	16.5 %	17.2 %
CTC-RNN	1e-7	12.8 %	15.4 %	15.2 %
CTC-RNN	1e-8	**11.9 %**	**14.5 %**	**14.4 %**

Unfortunately, an evaluation of the oral history test set on the CTC-RNN and the pDNN-HMM models was not possible. These were the subject of ongoing work by other researchers, and the oral history test set was not finalized at that time. Instead, we then evaluated the oral history set on the most recent model in Audio Mining at that time. This baseline system yields a 55 % word error rate on the oral history test set. We use this error rate and the results on the broadcast sets as a baseline for comparison of the first experiments.

Since the other acoustic models in Table 3.6 achieve improved recognition accuracy for broadcast data, it is reasonable to assume that they would also perform better on oral history than the baseline system. However, as the experiments in Chapter 4 show, with LF-MMI models—especially with the proposed data augmentation—we achieve the results that consistently and substantially outperform the benchmarks from Table 3.6—even when only one-eighth of the training data is used.

3.5 Preliminary Investigation of the Influence of Language Models

In Section 2.1.2 of the fundamentals chapter, we presented how speech recognition systems based on hidden Markov models can be decomposed into three independent models: the acoustic model, the language model, and the phonetic pronunciation lexicon. Later on, in the systematic review in Section 3.3.2, we structured the different challenges of oral history interviews by grouping them into four categories of challenges and stated whether the acoustic or language model models the category. Based on these conclusions, we have already elaborated that improving the

acoustic model is essential in the presented research for improving speech recognition on oral history interviews. In the following, we will additionally examine the influence of the language model in advance using intrinsic evaluation metrics to assess to what extent an improvement of this model is necessary. Based on these results, we conclude by arguing why the two broadcast language models we apply for the acoustic model experiments in this work can be reasonably applied for speech recognition of German oral history interviews—although generally being out-of-domain for this task.

3.5.1 Overview

We use an established language model trained for the conventional broadcast use case of the Fraunhofer Audio Mining for all experiments in the present research work. This model has already been used in the previous work by Stadtschnitzer [2018] for the long-term development of the German broadcast Audio Mining system. This language model has been trained using the *IRSTLM language model toolkit* [Federico et al., 2008]—as all language models in this work have been.

In contrast to the model used by Stadtschnitzer, we apply a pruning threshold of 1e-8 due to the improved results with this setup in Table 3.6. We refer to this model as the *default language model* or *default LM* in this work. Between 2018 and 2019, the default language model was being replaced in Audio Mining by a more recent model trained on significantly more text data—as can be derived from Table 3.7. For research, this *large language model* was first introduced in [Gref et al., 2020], cf. Chapter 6, and was used for additional, independent decodings in addition to the default language model. We perform these independent decodings to ensure the acoustic model's improvements using our proposed approaches are consistent and do not depend on a certain language model. The performance of both language models in terms of the word error rate is evaluated in Section 4.4.5 with multiple acoustic models.

The phonetic pronunciations of the decoding language model vocabulary and the acoustic model training vocabulary are obtained using the same grapheme-to-phoneme pronunciation model. This model is trained with the Sequitur G2P toolkit [Bisani and Ney, 2008] using the German pronunciation dictionary *Phonolex*[5] from the *Bavarian Archive for Speech Signals* (BAS), cf. Schiel [1998].

In the first acoustic model experiments of the presented work, the language model weight (*LMWT*) (cf. Equation A.1) was individually obtained for each model by decoding on a *broadcast development set* and fixed for that model. This development set is a subset of the GerTV1000h corpus, cf. Stadtschnitzer [2018, pp. 45 ff.]. However, since the set stems from the broadcast domain, this approach

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

Table 3.7: Overview of the two primary language models used throughout the presented research work. If not stated otherwise, *Default LM* is used as the decoding language model for the experiments. *Large LM* was published in the last phase of the presented research work and used for additional decoding in these experiments.

	Default LM	**Large LM**
Domain	Broadcast/News	Broadcast/News
N-Gram	5-gram	5-gram
Pruning Threshold	1e-8	1e-8
Vocabulary size	506,858	2,047,124
Running words in training data	76 million	1.6 billion
Sentences in training data	5.2 million	8.9 million

was considered not representative in later experiments. Therefore, we then moved on to apply the same predefined language model weight for all models in all experiments and did not further optimize this decoding parameter. All other decoding parameters are also kept equal for all experiments.

3.5.2 Perplexity

The *perplexity* is an *intrinsic evaluation metric*, cf. Jurafsky and Martin [2009, p. 95], for evaluation of the quality of language models independent from its application—such as in speech recognition. The perplexity $\mathrm{PP}(\boldsymbol{w})$ of a language model on given word sequence $\boldsymbol{w} = (w_1, ..., w_n)$ can be defined, cf. Jurafsky and Martin [2009, pp. 95–97], as the inverse probability of the word sequence, normalized by the number of words:

$$\mathrm{PP}(\boldsymbol{w}) := \sqrt[n]{\frac{1}{P(\boldsymbol{w})}} = \sqrt[n]{\frac{1}{P(w_1, ..., w_n)}}.$$

For N-gram language models, as they are commonly used in speech recognition, the overall probability of a word sequence is the product of all conditional probabilities of the recent word k given the history of the previous $N-1$ words, cf. Jurafsky and Martin [2009, p. 86–88]. Thus, for a bi-gram language model ($N = 2$), the perplexity is

$$\mathrm{PP}(\boldsymbol{w}) = \sqrt[n]{\frac{1}{\prod_{k=1}^{n} P(w_k|w_{k-1})}}. \tag{3.2}$$

Table 3.8: Perplexity of the two different language models used throughout the presented research on several evaluation sets.

Data Set	Default LM	Large LM
DiSCo Planned Clean	715.13	690.75
DiSCo Planned Mix	847.76	728.54
DiSCo Spontaneous Clean	494.24	576.27
DiSCo Spontaneous Mix	575.20	628.23
German Broadcast 2016	639.68	524.23
Challenging Broadcast	1152.90	946.13
Oral History	769.56	767.27
Interaction (Linguistics)	1480.74	1748.00
Spoken QALD-7	1626.62	1097.35

To better assess the perplexity of a given language model, a zero-gram language model with the same vocabulary can be considered for comparison. Zero-grams represent the most simple case of n-grams (with no prior knowledge) and have a uniformly distributed probability for all words:

$$P_{\text{0-gram}}(w_k) = \frac{1}{V}$$

with V as the vocabulary size of the lexicon. Applied to Equation 3.2, it becomes evident that the perplexity of a zero-gram language model is equal to the vocabulary size V. Thus, the perplexity of a given language model indicates how much better it models the test set than a simple zero-gram.

In speech recognition, the perplexity of a language model on a given data set is often used to estimate the quality of a language model in advance. This is based on the assumption that a low perplexity of language models is correlated with a lower word error rate of a speech recognition system using the respective language model. This correlation has been theoretically and experimentally studied and confirmed, among others, by Klakow and Peters [2002].

We investigate the usability of the two broadcast language models for the oral history use case in terms of perplexity. We compare the perplexity of the language models on broadcast data with that on other domains in Table 3.8. For both the default and the large language model, the perplexity on oral history is in a similar value range as the different broadcast sets. The perplexity of the models on oral history is slightly higher than for the DiSCo sets but much lower than for Challenging Broadcast. While the perplexity on the test sets of the other domains—namely interaction and question answering—is the highest overall in

Table 3.9: Out of vocabulary rates of the two different language models used throughout the presented research on several evaluation sets.

	Running OOV Rate		Unique OOV Rate	
Language Model:	**Default**	**Large**	**Default**	**Large**
DiSCo Planned Clean	0.75 %	0.30 %	2.25 %	0.95 %
DiSCo Planned Mix	1.20 %	0.46 %	3.58 %	1.30 %
DiSCo Spontaneous Clean	0.44 %	0.21 %	2.22 %	1.07 %
DiSCo Spontaneous Mix	0.70 %	0.22 %	2.61 %	0.87 %
German Broadcast 2016	1.39 %	0.45 %	3.58 %	1.43 %
Challenging Broadcast	2.27 %	0.82 %	6.94 %	2.68 %
Oral History	1.72 %	0.96 %	8.43 %	4.63 %
Interaction (Linguistics)	1.46 %	0.72 %	6.18 %	3.31 %
Spoken QALD-7	1.57 %	0.68 %	3.69 %	1.61 %

the table, this is not true for oral history. Thus, we conclude that the broadcast language models do not pose a significant disadvantage for speech recognition on oral history compared to broadcast.

3.5.3 Out-of-Vocabulary Rates

The *out-of-vocabulary rate* (*OOV rate*) is another measure that can be used for the prior assessment of the quality of language models—in particular, the vocabulary of a language model. The OOV rate is the ratio of words in a test set that do not appear in the pronunciation lexicon to all words in the test set. Words that do not appear in the lexicon and ultimately cannot be recognized by conventional speech recognition systems. Thus, high OOV rates can indicate poor performance of a language model on a test set in advance.

Two variants of OOV rate are common, cf. Stadtschnitzer [2018]: the *running OOV rate* where the total number of occurrences of each (*running*) word is counted, and the *unique OOV rate* (or *vocabulary OOV rate*) where each different word is counted only once. The running OOV rate is of particular relevance for the overall performance of a language model in ASR on a given test set since it can be interpreted as a lower bound for the word error rate for word-based systems.

The running OOV rates of the default language model range from 0.4 % to 2.3 % on the different test sets, as shown in Table 3.9. The rate is the lowest for the DiSCo sets and the highest on Challenging Broadcast. The running OOV rates on the test sets from the other domains all fall in this range of values of the broadcast

data—and are only slightly higher than on German Broadcast 2016. This indicates that out-of-vocabulary is not a major issue of the default language models for these domains.

In general, the OOV rates of the large broadcast language model are consistently lower than of the default model. The running OOV rate of the large model is below one percent on every data set and often even below 0.5 %. Compared to the default language model, the relation of the OOV rates on oral history and the broadcast test sets shifts with the large model. The large model has the highest, however, overall fairly low running OOV rate on oral history with 0.96 %.

For both language models, the unique OOV rate on oral history is, by far, the highest overall. However, this is not reflected in the running OOV rate, suggesting that there are many words in the vocabulary of oral history that do not appear in the vocabulary of the language models and only seldom appear in the running text of the oral history set. This is not surprising considering the very diverse, often historical, subjects of the interviews. Since these words occur rarely, they do not pose an issue for evaluating the acoustic model via the word error rate using the broadcast language models. For the final application in productive applications such as Audio Mining, however, it may be useful or even necessary to apply an oral-history-specific language model to enable the recognition of these words.

It should be noted that the appearance of a word in the vocabulary of the language model alone is not sufficient for it to be recognized by a speech recognition system. It must also appear sufficiently often in the text data used for training data the N-grams—which is ultimately reflected in the perplexity of the language model. Therefore, both measures should always be considered jointly. Furthermore, the phonetic pronunciation in the lexicon must correspond to the actual pronunciation of the speakers to be recognized by the ASR system.

3.5.4 Conclusion

Overall, both the perplexity and the OOV rate evaluation on oral history achieve similar or slightly higher values than the sets from the broadcast control group. This is true for both broadcast language models in question and indicates that the broadcast models can be reasonably applied for speech recognition of German oral history interviews. Thus, the language model does not require prioritized improvement in the presented research work. However, this does not mean that further improvement of recognition accuracy would not be possible and valuable with a domain-specific language model for oral history. Thus, we focus on the acoustic model for further research since we conclude from the previous analyses that this is the essential component to be improved.

3.6 Human Word Error Rate Estimation for Oral History Interviews

We have highlighted the limitations of the word error rate as an evaluation metric for speech recognition in Section 3.4.13. In the following section, we present our approach to computing a human word error rate for German oral history. This study aims to put the recognition performance of automatic systems in perspective to human transcriptions and to uncover challenges in oral interview transcription. We present the pipeline for our study, the results, and conclude with a discussion of transcription errors.

3.6.1 Related Work

To the best of our knowledge, no work estimated a human word error rate for oral history interviews so far—particularly for German Oral History data. However, such a measure has been estimated for individual data sets. Lippmann [1997] performed one of the early works comparing transcripts of speech recognition systems with human word error rates. The authors estimate the human word error rate for different domains using common English corpora, such as the WSJ for read-speech and Switchboard for conversational speech. The authors report a human error rate of about 1 % for the Wallstreet Journal Corpus (WSJ) [Paul and Baker, 1992] and 4 % for Switchboard [Godfrey et al., 1992].

With the enormously increased recognition performance of ASR in the last decade, Xiong et al. [2017a] and Saon et al. [2017] reconsidered the human word error rate on the English Switchboard corpus. The reported human error rate of these works was in the range of 5.1—5.9 %. While proposed as an error rate for this particular corpus, this human word error rate was sometimes misconstrued as a general human word error rate in the general public. A detailed overview of Xiong et al.'s and Saon et al.'s approaches for human word error rate estimation is given in the following, where we present and compare our approach for the estimation.

3.6.2 Annotation Approach and Experimental Setup

In the following experiment, we estimate a *human word error rate* on transcribing German oral history interviews, inspired by the experiments of Xiong et al. [2017a] and Saon et al. [2017]. Strictly speaking, this human error rate is the difference of transcriptions between two transcribers. One transcriber is taken as the reference and the other as the hypothesis. These results aim to expose the challenges even humans face transcribing oral history interviews and put the achieved error rates of speech recognition systems in this domain into perspective.

Xiong et al. [2017a] used a two-staged transcription pipeline by a large commercial vendor to transcribe the English Switchboard data for their experiments. In the first stage, one professional transcriber annotates the pre-segmented speech from scratch. A second transcriber corrects the first transcriber's transcription in the second stage. Using this approach, the authors report a 5.9 % human error on the English Switchboard data.

Later on, Saon et al. [2017] replicated the experiment on the same data as the authors consider the values reported by Xiong et al. to be too high. The authors also used a two-staged pipeline by a large commercial vendor, where three different transcribers transcribed the speech segments from scratch independently from each other. Then, in the second stage, a fourth transcriber performed a quality check and corrected the annotations of the first stage. Saon et al. report a human word error rate in the range of 5.1–5.6 % on English Switchboard.

We could not apply such a professional commercial pipeline with the project's budget. Instead, we aim at approximating the two-stage procedure by replacing the first stage with our Audio Mining speech recognition system and letting humans correct the raw ASR transcript. For this system, the adapted oral history acoustic model from Section 5.3 was applied. For the language model, broadcast data was combined with manual oral history transcripts in roughly equal proportions.

We use the ASR result to chunk the interviews into short segments at the longest speech pauses until we obtain segments of 30 seconds or less. We obtain 3,122 segments for the 10 hours of data by this approach. Thus, the average segment length for our data is 11.5 seconds. In the second stage, three human transcribers were independently provided with the same raw ASR transcript and were asked to correct it.

Our experiment was performed near the end of the presented research work in 2021-2022 in a joint project with the *Haus der Geschichte* (HdG) foundation. For our experiment, 10 hours of German oral history interviews from the *Zeitzeugenportal* of the *Haus der Geschichte* foundation were used (cf. Section 3.4.10). The HdG portal's interviews are somewhat different from the *Deutsches Gedächtnis* data, primarily studied in the presented research work. The interviews are recorded with more professional equipment, thus having better audio quality. In most other aspects, however, the data is very similar, so we assume that the conclusions of the experiments can also be applied to the other data.

The transcription was performed by three employees at the *Haus der Geschichte*, who have an academic background in history—but are not professional transcribers. The transcribers did not only correct the ASR transcription but also annotated the perceived emotions and sentiment for each segment. We use these annotations to further study the influence of emotions and sentiments in oral history.

Table 3.10: Overview of the human word error rates in percent between two different transcribers (Tr.): A, B, and C. Three different types of experiments were performed to investigate different reasons for the resulting error rates.

	Setup	Reference	Hypothesis Tr. A	Hypothesis Tr. B	Hypothesis Tr. C
1)	Case-sensitive WER Hesitations counted as word errors	Tr. A	-	7.77	9.51
		Tr. B	7.79	-	8.83
		Tr. C	9.54	8.83	-
2)	Case-sensitive WER Hesitations ignored	Tr. A	-	7.41	9.00
		Tr. B	7.45	-	8.58
		Tr. C	9.06	8.58	-
3)	Case-insensitive WER Hesitations ignored	Tr. A	-	6.54	8.43
		Tr. B	6.58	-	7.74
		Tr. C	8.48	7.75	-

3.6.3 Results

The results of comparing the different human transcriptions are summarized in Table 3.10. We compared three different setups, one after the other, to investigate different causes for the error rates.

We begin with Setup 1 in the top third of Table 3.10 that we usually also consider when evaluating speech recognition systems: The word error rate is calculated case-sensitively, i.e., different casings of the same words are counted as substitutions in the word error rate. Furthermore, transcribers were asked to transcribe hesitation sounds with a predefined spelling. Our ASR system usually transcribes these hesitation sounds if they can be heard clearly enough. The highest difference is between transcriber A and C, with a 9.5 % word error rate with this setup. The lowest is between transcriber A and B, with a 7.8 % word error rate.

For a more detailed analysis, we first consider the combination with the lowest error rate in this setup: Transcriber A as the reference and Transcriber B as the hypothesis. Transcriber A has overall 78,428 transcribed words that are used as the reference. Comparing Transcriber B to A, Transcriber B has 6093 errors—1106 insertions, 1328 deletions, and 3659 substitutions. An overview of the top five errors for each category is given in Table 3.11. The most common differences (*errors* in terms of the WER) are insertions of the hesitation sounds *äh* (German variant of the hesitation *er* or *uh*), *Äh*, and *hm*, by Transcriber B. It seems Transcriber B was paying more attention to hesitation sounds than Transcriber A. However, it is

Table 3.11: Top five errors for each error type between Transcriber A and B for a case-sensitive human word error rate estimation on German oral history data where hesitation sounds are annotated and counted as word errors.

Error Type	Transcriber A (Reference)	Transcriber B (Hypothesis)	Error Count
Deletion	und	-	71
Deletion	ja	-	63
Deletion	ich	-	47
Deletion	dann	-	37
Deletion	in	-	37
Insertion	-	äh	118
Insertion	-	hm	50
Insertion	-	Äh	38
Insertion	-	und	35
Insertion	-	die	25
Substitution	habe	hab	61
Substitution	sie	Sie	61
Substitution	dass	das	43
Substitution	ich	Ich	39
Substitution	und	Und	39

noteworthy that both transcribers have the same annotation of hesitation sounds way more often than not. Both transcribed the most common hesitation sound *äh* at the same position in 675 cases. In 159 cases the annotation of *äh* differed and led to an error.

The next common error type comparing Transcriber A and B is deletions of short words—*und* (*and*) and *ja* (*yes*)—that Transcriber A has annotated quite often, but Transcriber B has not. The next most common errors are substitutions of the same words in slightly different spellings: e.g., formal *habe* vs. informal *hab* (*have*), and casing errors. These observations lead us to two questions that will be investigated with two further setups: what is the influence of hesitation sounds on the error rate? And what is the influence of the casing?

To answer the first question, we remove the hesitation sounds from the transcript of all transcribers and compare the transcripts again. The resulting error rates are depicted as Setup 2 in the middle part of Table 3.10. Overall, without hesitation sounds, the word error rate decreases by 0.3–0.5 percentage points. Since the sounds were removed from both the reference and hypothesis, the overall influence on the human error rate is quite limited.

Table 3.12: Word error count comparison between Transcriber A (ref) and Transcriber B (hyp) for the three different human error rate investigation setups.

Setup	WER	Errors	Num. Ref	Ins.	Del.	Sub.
1)	7.77 %	6093	78,428	1106	1328	3659
2)	7.41 %	5733	77,383	835	1288	3610
3)	6.54 %	5059	77,383	835	1288	2936

Comparing Transcriber A with B again, the total number of errors decreases by 360 errors by removing hesitation sounds, as reported in Table 3.12. The largest share of this is accounted for by decreasing insertions, which may already be assumed in advance. However, at the same time, the total number of words in the reference decrease even more by more than 1000 words.

Lastly, we consider the second question formulated earlier and examine the influence of casing errors. For this purpose, we compare the transcriptions with removed hesitations and, again, without taking the casing into account. The results are depicted as Setup 3 in the bottom third of Table 3.10. Ignoring the casing for evaluation reduces the word error rate by a further 0.5 to just under 1.0 percentage points. As shown in Table 3.12, ignoring the casing naturally reduces only the number of substitutions.

As shown in Table 3.13, after removing the hesitation transcriptions and lowercasing all words, the top five inserted words by Transcriber B are now also mostly short words with only one syllable—words that can be easily overheard in spontaneous speech, especially when there are word repetitions or ungrammatical sentences due to rephrasing. However, the top five errors per category account for just under 11 % of all word errors with this setup. Therefore, a large share of the errors is distributed among many individual errors that are not as systematic as these.

Finally, we take the arithmetic mean of the six different transcriber pairs for each setup to report a human word error rate for each of the three different analysis scenarios we studied. These values are given in Table 3.14, in addition to the standard deviation.

For the evaluation we perform in our research—case-sensitive word error rate evaluation and annotating hesitations—the corresponding human word error rate on oral history interviews is 8.7 %. There are two primary reasons why we evaluate our ASR with Setup 1. For the indexing of the content and adequate readability, the casing of words in the German language is crucial and should be correctly transcribed by systems. In our ASR system, the casing is part of the language model and pronunciation lexicon since we achieve better overall recognition results

Table 3.13: Top five errors for each error type between Transcriber A and B for a case-insensitive human word error rate estimation with ignored hesitations on German oral history data.

Error Type	Transcriber A (Reference)	Transcriber B (Hypothesis)	Error Count
deletion	und	-	73
deletion	ja	-	64
deletion	ich	-	47
deletion	dann	-	39
deletion	in	-	38
insertion	-	und	35
insertion	-	die	26
insertion	-	da	24
insertion	-	ich	22
insertion	-	dann	20
substitution	habe	hab	62
substitution	dass	das	43
substitution	das	dass	35
substitution	dann	da	18
substitution	das	es	18

than with a downstream inverse-text-normalization component. Additionally, as described in Section 3.4.13, transcribed hesitations are crucial in specific research questions for oral history. They help historians and other researchers assess not only what but how something was said in an interview. Therefore, we also evaluate these hesitations in our ASR evaluation.

3.6.4 Discussion and Limitations

Compared to the 4.0—5.9 % human word error reported for English Switchboard, the human error rates on German oral history data we obtained are significantly higher. This is particularly because of the characteristics of oral history interviews, which were pointed out and discussed in Section 3.3.2.

At the same time, we must admit that the experiment is subject to some limitations. The calculated error rate depends on the transcribers, their motivation, and the applied procedure. On the one hand, presumably, it would be possible to reduce the differences between the transcribers by several correction iterations. On the other hand, we find it remarkable that all transcribers had the same raw

Table 3.14: Average human word error rates on 10 hours of manually transcribed German oral history interviews. The average human word error rate is given as the arithmetic mean ± the standard deviation of the six different comparisons in each setup.

Variant of Analysis	Avg. Human WER
Case-sensitive WER, including hesitations	8.71 % ± 0.79 %
Case-sensitive WER, excluding hesitations	8.35 % ± 0.74 %
Case-insensitive WER, excluding hesitations	7.59 % ± 0.86 %

ASR transcript as a basis—and yet such comparatively large differences in the annotations can be found. It can be assumed that applying an annotation from scratch would result in an even higher error rate.

Another advantage that humans have over the ASR system is the context. The transcribers were aware of the content discussed in the interviews and could listen to previous and subsequent segments at will. Our ASR system does not have this advantage and transcribes segment by segment independently. For a fair comparison, the ASR system would need to account for surrounding segments—which is currently not supported for standard n-gram language models.

If the human transcribers had listened and annotated each segment in a random order, this would naturally result in a higher human word error rate. Finally, the transcribers were provided not only with the audio but also with the video stream. It is well-known that visual feedback, e.g., seeing the lip movement, can improve speech understanding. This can be another advantage for human transcription. Nevertheless, based on the annotator's feedback, transcribing only audio segments (especially if in random order) would have resulted in significantly reduced motivation, which in turn would have spuriously affected transcription quality.

Lastly, it should be emphasized that the HdG oral history interviews used for this experiment have a fairly high audio quality and are quite easy to understand. This is not true to the same extent for many other oral history interview archives in different languages. The interviews in the default oral history test set from the archive *Deutsches Gedächtnis*, which we examine in most experiments in the presented research, have much more challenging acoustic conditions. For these interviews, a significantly higher human error rate can be assumed. We also observe this effect with ASR systems, which, despite robust training, often struggle with interviews with poor audio quality, as discussed in Section 5.5.

3.6.5 Conclusions

In this experiment, we investigated the accuracy of human transcription of German oral history data by comparing corrected versions of raw ASR transcription of three different persons. We estimate a word error rate of 8.7 % for recent oral history interviews with relatively clean acoustic conditions. We discussed the different types of possible human transcription errors of oral history interviews. This error rate is intended to serve us as a rough benchmark for estimating human transcription accuracy and is by no means to be taken as an absolute benchmark of human performance. We have discussed the limitations of our approach and argued that different approaches to estimating the human error rate can lead to different results—as has been the case, for example, with several works on the human error rate estimation on the Switchboard corpus in the past. We suspect our human error rate for oral history is more likely to be on the low end and may increase when transcribing from scratch or using a random order of segments. However, we think that our error rate estimate can serve as a reference when assessing ASR systems on oral history—and what realistic word error rates of ASR systems can be expected in the future.

3.7 Summary and Contributions

3.7.1 Summary

This chapter presents the oral history speech recognition use case that we study in the presented research work. In this chapter, we conducted preliminary experiments and investigations to assess the influence of the different components of speech recognition systems and different corpora used for the research questions.

First, we introduced the research project KA^3, in which large parts of the research work were carried out, and the project partner archive *Deutsches Gedächtnis*. The archive provided the oral history interviews to be studied for this research. We also motivated the work by pointing out the value of transcripts and automatic transcription systems, such as the Fraunhofer IAIS Audio Mining system, for oral history research. Performing a systematic (non-statistical meta) review of related works, we exposed the research gap in this field. We summarized and categorized the several challenges found by other researchers in their attempt to transcribe oral history interviews with ASR.

For our research, we introduced the different ASR corpora used in the experiments in the following chapters. In particular, we have presented and proposed the *German Oral History* test set, which was developed as part of this work to investigate the research questions. At the same time, we explained why we assume that it is not sufficient to investigate proposed systems that shall be applied

in real-world applications only on one benchmark—as it is usually done in ASR research. To assess the usability of a real-world application, an evaluation using several different data sets from different domains is mandatory. This coincides with recent works in ASR research, which also criticize standard ASR evaluation when it comes to real-world applications. To compare the similarities and differences of the corpora from diverse domains, we performed an initial statistical analysis of different attributes. We further studied the speech tempo of oral history interviews compared to other domains in terms of the phone rate. Although indicated in the literature, we found that a high speaking rate or speaking rate variance does not apply to our studied oral history interviews.

In another investigation, we studied the influence of language models trained on large news-text corpora for the automatic transcription of oral history interviews. Using intrinsic evaluation measures, such as perplexity and OOV rates, we demonstrated that, in particular for oral history, these language models are likely to be sufficiently well suited for application in speech recognition systems. This is quite surprising for two reasons. First, the news text used for training can be assumed to be quite out-of-domain for oral history. Second, the presumed performance with broadcast language models is significantly worse for other domains studied in our investigation, such as interaction or question-answering. These experiments confirm the approach in the following chapters to prioritize the acoustic model as the component to be improved in the speech recognition of German oral history interviews.

The chapter concludes with an experiment designed to demonstrate that transcribing German oral history interviews is challenging—even for humans. Using different comparisons of three transcribers, we present an 8.7 % estimated *human word error rate* on a ten-hour German oral history data set with clean acoustic recording conditions.

3.7.2 List of Contributions

List of scientific contributions in this chapter:

- A human word error rate on German oral history interviews in clean acoustic conditions was experimentally estimated.
- A systematic (non-statistical meta) review of related works on the several challenges of oral history interviews for automatic speech recognition was provided.
- A representative German oral history data set for ASR was created and proposed to evaluate speech recognition systems on oral history interviews.

- Several preliminary studies were conducted for ASR speech corpora and test sets from several domains. The insights inferred help to better assess and evaluate the proposed methods in the following chapters in terms of impact on different domains and domain overfitting. In particular:
  - Statistical analysis of the test sets was performed, exposing similarities and differences in the various speech domains.
  - The speech rate for the different test sets was estimated using a GMM-HMM-based phone rate estimation.
  - A preliminary analysis of two broadcast language models for oral history ASR and the other domains was performed using intrinsic evaluation metrics.

4 Acoustic Robustness for Oral History Interviews

In the previous chapter, we have found that although the baseline Audio Mining system of 2016/2017 delivers adequate results for broadcast data, the transcripts obtained for oral history interviews tend to be of poor quality. We discussed several challenges of oral history interviews for speech recognition systems, identifying the audio signal quality as one of the main challenges. Oral history interviews are usually recorded in the living rooms of the contemporary witnesses using commonly available recording equipment. This equipment changed over the years and resulted in a wide range of recording qualities. In particular, recent interviews often have a very decent recording quality. Other interview recordings suffer from intense background noises, reverberation, and additional unspecific distortions. Other significant challenges posed by oral history interviews are the colloquial language used in spontaneous speech, hesitations, age- and health-related changes in speaking, and domain-specific words used in the interviews that usually do not occur in everyday speech.

In this chapter, we focus on the acoustic distortions of oral history interviews by improving the robustness of acoustic models. As discussed in the previous chapter, avoiding domain overfitting is essential for the real-world application of speech recognition systems. This is particularly important for our task since the oral history interviews in our collection have a vast range of conditions. We have to expect further unseen conditions in interviews for future applications. Therefore, we evaluate and study all models, in this and all following experiments, on several test sets from different domains to obtain a reliable estimate of the real-world performance for seen and unseen conditions.

In this chapter, we proceed in several subsequent steps to obtain robust models. In Section 4.2, we first give an overview of the basic techniques for robust speech recognition for recent DNN-HMM acoustic models. Three sequential studies follow that all comprise several experiments.

The first study in Section 4.3 investigates and compares different acoustic models proposed in the current literature. In the second study in Section 4.4, we build upon the first study's results and explore multi-condition training to improve the acoustic robustness and narrow the domain mismatch between the training domain and out-of-domain data. We propose the combination of noise and reverberation data augmentation as a suitable approach to improve overall performance and

substantially improve speech recognition performance for oral history interviews. In the third study in Section 4.5, we additionally study speech enhancement to improve the acoustic robustness of oral history interviews. We compare different speech enhancement approaches for noise reduction and dereverberation for the oral history use case and compare them to the multi-condition approach. In Section 4.6, we summarize the chapter's findings and contributions.

4.1 Thesis Author Contribution

Parts of this chapter are covered in the publications:

> Michael Gref, Joachim Köhler, and Almut Leh. Improved transcription and indexing of oral history interviews for digital humanities research. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 3124–3131. European Language Resources Association (ELRA), 2018a. URL https://aclanthology.org/L18-1493

> Michael Gref, Christoph Schmidt, and Joachim Köhler. Improving robust speech recognition for German oral history interviews using multi-condition training. In *13th ITG Conference on Speech Communication*, pages 256–260. VDE / IEEE, 2018b. URL https://ieeexplore.ieee.org/document/8578034

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author.

In the present chapter, we summarize and extend the experiments in [Gref et al., 2018a] on CE-LSTM models trained on 128 hours of speech and put them into perspective to the experiments of [Gref et al., 2018b] on LF-MMI–TDNN-LSTM models trained on 1000 hours of speech, presented in Section 4.3 and Section 4.4. In the presented research work, we also performed additional experiments and analyses on the models trained in the aforementioned papers, combining and extending the findings, inferring further insights, and putting the results into perspective to our experiments in the following chapters.

4.2 Overview of Robust ASR Approaches

The oral history interviews examined in this work were recorded in the past, often many years or even decades ago. Evidently, the recording conditions cannot be changed in retrospect. At that time, digital processing of this data was often not even conceivable, so the main focus was on making the recordings intelligible

to humans. However, these recording conditions determine not only the resulting recording quality but also the selection of the algorithms to be considered for robust speech recognition. For most oral history recordings in question, only one microphone was used so that multi-channel signal processing approaches, like acoustic beam-forming or multi-stream ASR, cf. Li et al. [2016, pp. 239 ff.], are not feasible. Thus, we only consider single-channel recording approaches for our further work.

A wide variety of approaches for single-channel robust speech recognition have been studied in the past. An overview of these approaches, with the main focus on noise-robustness, is given by Li et al. [2014]. Many of these approaches are particularly used for GMM-HMM systems. Some of these approaches—such as CMN/CMVN, feature space transformations, and speaker adaptive training (cf. Section 2.1.3)—are applied by default in the bootstrap training for the experiments in this work. The bootstrapping is described in Section 4.3.1.

Several categorizations of robust ASR based on different attributes of the approaches can be found in the literature. A common categorization is *feature-domain* vs. *model-domain* approaches, cf. Li et al. [2014]. The *feature-domain* approaches either try to utilize features that are inherently robust to the distortions or domain, or to transform the features at inference time to match the feature distribution from the training domain without any changes of the model's parameters. Thus, feature-domain approaches can be subdivided into *speech enhancement* and *robust feature* approaches, cf. Josifovski [2002, pp. 8–31]. The *model-domain* approaches, on the other hand, modify the model's parameter to obtain a model that is robust to the distortions.

In the days of GMM-HMM speech recognition systems, a wide range of methods were combined to achieve robustness. However, in the era of hybrid ASR systems, more straightforward methods that mainly rely on the powerful modeling capability of deep neural networks seem to prevail. Speech enhancement approaches are often applied to improve the speech signal quality in advance—e.g., by reducing noises and compensating channel distortions—aiming to obtain a speech signal with only slight distortions that match the training conditions. For the model-domain side, *multi-condition* (or *multi-style*) training is commonly applied. It aims at training the acoustic model on a wide range of conditions so that the model learns to generalize and rely primarily on robust speech features in given signals, cf. [Lippmann et al., 1987].

Multi-condition training and speech enhancement for robust ASR using hybrid systems was studied, among others, by Tang et al. [2018]. The authors compared different approaches in terms of a domain adaptation problem for distant speech recognition using a hybrid TDNN acoustic model. For their experiments, the authors found that multi-condition training achieves the best results among the four

studied approaches using the AMI corpus [Carletta et al., 2006]. This observation is also consistent with the results of one of our preliminary works [Hirsch and Gref, 2017] on the Aurora4 challenge [Parihar and Picone, 2002]. We also compared speech enhancement—both using the magnitude and phase STFT spectrum—and multi-condition training for a feed-forward hybrid ASR system using the Aurora4 corpus. We found that multi-condition training yields the best results among the studied approaches.

It is noteworthy that the aforementioned works studied comparatively restricted scenarios of acoustic distortions—e.g., distant speech only, artificially added noise with predefined signal-to-noise ratio only—and the type and nature of the distortion are known in advance. However, we face previously unknown combinations of different acoustic distortions in the oral history interviews. For many interviews, it can be assumed that noise and room reverberation occur simultaneously. The noise may have occurred in the room itself where the recording took place. However, the noise may also be due to the recording device itself—such as noises from cassette recorders commonly used a few decades ago. The noise or other distortions could also be unintentionally added years after the recording—such as deterioration of the magnetic tapes due to age, artifacts from multiple conversions, or digitization. Therefore, an open research question is which approaches can be generalized to these complex real-world problems and which are suitable for the problems at hand. The following studies on different approaches for the German oral history use case in this chapter aim to contribute to this research question.

4.3 Study: Comparison of Selected Hybrid Acoustic Models

Automatic speech recognition is a highly active research field. New model architectures and training approaches are proposed regularly. Although many of these models and approaches are promising to show good performance for the task at hand, it is not feasible to evaluate different robust speech recognition approaches with a multitude of models. This is due to the computational effort and time required to train hybrid neural network acoustic models with millions of parameters on close to 1000 hours of annotated speech. Thus, in the literature, as in many previously mentioned related works, one fixed model configuration is often chosen for the experiments.

At the same time, for a real-world application, it is necessary to determine which models generalize well for real-world data and which models suffer from domain overfitting. In this section, we want to experimentally determine which model we can expect to produce the best results for our real-world use case. We

compare selected acoustic models with different amounts of training data to make a statement about how well the models are able to generalize. In addition, we investigate the influence of segment lengths on different types of models since we observed differences in recognition performance in initial experiments. We will use the most promising model for subsequent experiments in the presented research work.

4.3.1 Experimental Setup

In the following, we describe the experimental setup we use for the model evaluation. We conducted our experiments with the Kaldi ASR toolkit [Povey et al., 2011] because it was the most promising ASR toolkit in research in 2017. Researchers developed new approaches and integrated them quickly into the open-source toolkit due to its increasing popularity, cf. Section A.1 in the appendix.

DNN-HMM Acoustic Models

The selection of models for our experiments is based on promising models from the literature that performed well on large-vocabulary continuous speech recognition tasks in 2015–2017. Primarily, results on the English Switchboard task Godfrey et al. [1992] were considered since it was one of the predominantly studied speech recognition challenges at the beginning of the presented work. In the following, we describe the different acoustic models we study.

- **Cross-Entropy (CE)** trained hybrid acoustic models as described in Section 2.2:
  - **CE–LSTM** is a cross-entropy-trained model with a common, unidirectional LSTM neural network architecture with three stacked LSTM layers, cf. Sak et al. [2014]. The *nnet3* implementation for neural networks in Kaldi we apply for this, and the other models, 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 layers have a cell dimension of 1024 and a projection dimension of 256.
  - **CE–BLSTM** is a cross-entropy-trained model similar to CE–LSTM but with bidirectional instead of unidirectional LSTM layers, cf. Chen and Huo [2016]; Zhang et al. [2016]; Zeyer et al. [2016]. The bidirectional LSTM layers have the same cell dimension as the unidirectional ones. However, the projection dimension is reduced to 128 to compensate for the increased computational load required for bidirectional back-propagation through time.

- **LF-MMI** trained models use the sequence discriminative training of neural network acoustic models based on the LF-MMI implementation [Povey et al., 2016], as described in Section 2.4.3. Due to the reduced frame rate, LF-MMI models are significantly faster to train than conventional cross-entropy models. Thus, LF-MMI makes it feasible to experiment with larger model architectures than cross-entropy-trained models.

  We use a TDNN-LSTM architecture combining *time delay neural network* (TDNN) with LSTM layers, cf. Cheng et al. [2017]; Peddinti et al. [2018]. Our studied TDNN-LSTM models have ten hidden layers in an architecture proposed and investigated by Cheng et al. [2017]. The models combine seven 1024-dimensional TDNN layers, using TDNN-subsampling [Peddinti et al., 2015b], and three 1024-dimensional LSTM layers stacked in the order given in Figure 4.1. The projection dimension of the LSTM layers is 256.

  Overall we compared two training routines for LF-MMI–TDNN-LSTM:

  - **LF-MMI–TDNN-LSTM (standard)**: A standard training configuration used in Kaldi for the English Switchboard ASR challenge.
  - **LF-MMI–TDNN-LSTM (per-frame-dropout)**: This setup extends the standard LF-MMI training configuration with the application of per-frame-dropout for LSTM layers, as proposed by Cheng et al. [2017]. The dropout schedule is illustrated in Figure 4.2. This training routine uses the (default) LSTM layer implementation *LSTMp* instead of *FastLSTMp* to apply the per-frame dropout on the LSTM layers. Furthermore, the factor for cross-entropy regularization is increased to 0.025 from 0.01.

All acoustic models in our experiments use the same 300-dimensional input at each time-step consisting of five consecutive 40-dimensional MFCC features and a 100-dimensional i-vector [Dehak et al., 2011] estimated in an *online fashion* for speaker adaptation.

The total number of parameters of the models slightly varies for each training configuration and data since the number of output nodes is always equal to the number of tied states obtained in the last HMM-training stage of bootstrap training. A maximum of 11,500 states (or *leaves* in the phonetic decision tree) is configured in the training routines we use for conventional, cross-entropy training. In practice, the number is usually smaller. In our experiments, it is usually between 9000 and 9500 states. For LF-MMI training, the number of states is generally further reduced, cf. Povey et al. [2016]. A maximum of 7000 states is configured in the training routines we use. In practice, the number is usually between 5000 and 6500 states.

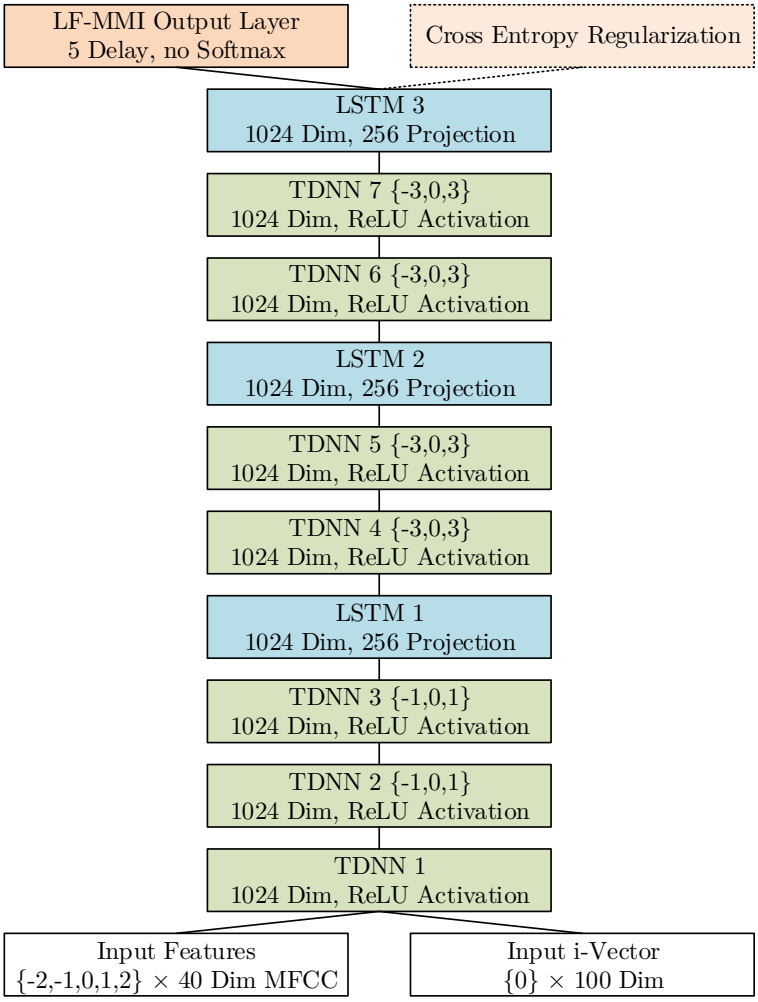


Figure 4.1: Neural network architecture of the LF-MMI trained LSTM-TDNN neural network.

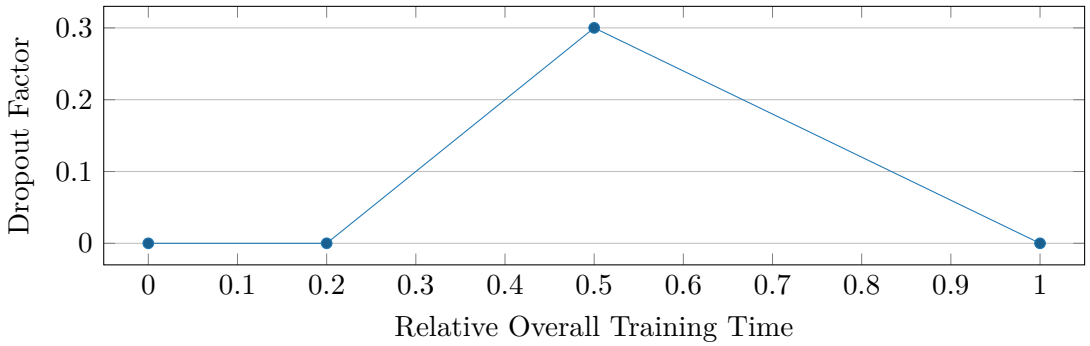


Figure 4.2: Per-frame dropout schedule for the LF-MMI acoustic model training. The schedule follows the general recommendation by Cheng et al. [2017].

Table 4.1: Comparison of the number of parameters of the studied models. We only report an approximate number of parameters for the total number of parameters since the exact number is subject to the number of tied states from the HMM-bootstrap training used as output nodes.

Model	Total Parameters	Hidden Parameters
CE–LSTM	15 Million	10,164,736
CE–BLSTM	8 Million	5,707,520
LF-MMI–TDNN-LSTM	39 Million	33,367,552

Since the total number of parameters of the models are subject to this variable number of tied states, the total number of parameters is only of limited use for comparison. Complicating matters, a particular property of LF-MMI training is cross-entropy regularization, implemented and trained as an additional, second output layer during training, as described in Section 2.4.3. This layer additionally increases the total number of training parameters for LF-MMI even though it is not used during decoding.

However, additionally to the total number of parameters, the number of parameters of the hidden layers can be used to compare the model. These values are given in Table 4.1. As to be expected, the BLSTM model is the smallest one. This is due to the reduced projection dimension, as described before. The TDNN-LSTM model is significantly larger than the other models.

We explicitly point out that the choice of models for the experimental setup is unsuitable for stating which network architecture or training approach is generally better for ASR. We compare differently sized neural networks, respectively, with the two different training approaches. However, this is explicitly not the objective of this study and has been extensively investigated in other works, e.g., by Povey et al. [2016]. LF-MMI allows the training of large models with LSTM layers on large-scale ASR corpora in a reasonable amount of time—which is not the case to the same extent for cross-entropy models. We take advantage of this to find the most promising acoustic model variant for further experiments.

Bootstrap Training

The cross-entropy and LF-MMI training rely on GMM-HMM models from bootstrap training for time alignment and state-tying. Since the GerTV1000h corpus does not contain any phonetic alignments, we perform *flat-start* bootstrap HMM-training with Kaldi. We start with a simple model to obtain initial alignments that are subsequently improved by more complex HMM models.

Table 4.2: Configuration of GMM-HMM model training in bootstrap for DNN-HMM acoustic model training in the Kaldi model comparison experiments. The models are trained subsequently from top to bottom. Each model is used to time-align the training data for the next training stage.

Model	Training Data	Features	Tied States
Monophone	11,000 shortest utt.	$\Delta\Delta$ + CMN	252
Triphone 1	38,000 rand. utt.	$\Delta\Delta$ + CMN	$< 3{,}200$
Triphone 2	38,000 rand. utt.	$\Delta\Delta$ + CMN	$< 4{,}000$
Triphone 3	entire data	LDA+MLLT	$< 6{,}000$
Triphone 4	entire data	LDA+MLLT+FMLLR	$< 11{,}500$

The bootstrapping configuration for the present experiments is shown in Table 4.2. First, we train a simple monophone GMM-HMM on the shortest utterances of the training data to facilitate initial alignment. We use 13-dimensional MFCCs with delta-delta features and cepstral mean normalization, cf. Section 2.1.3, for this and the subsequent two models. As presented in Table 4.2, four triphone GMM-HMM models are trained after the monophone model on increasing training data. Each model is used to generate the alignments for the training of the subsequent model.

This bootstrap training configuration is kept almost equal for all following experiments in the presented research work, except for the amount of training data. For the present experiment, the amount of training data is kept fixed and relatively small to keep experiments with different amounts of training data in the DNN-HMM training stage comparable. In later experiments, where we use the entire training data, we increase the training data for the bootstrap models. The influence of increased training data in bootstrap training on the final acoustic model is studied in Section 4.4.4.

Training Data

At the beginning of the presented research work, we do not have representative training data available for the German Oral History task at hand. Therefore, we are interested in how well the studied models generalize with different amounts of training data. Consequently, we conduct two experiments for the different models: one training using the entire GerTV1000h training data set, cf. Section 3.4.2, and one using about one-eighth of the training data. For this subset, we use precisely 100,000 randomly chosen utterances from the GerTV1000h corpus, resulting in a training subset of 128-hour length.

Table 4.3: Comparison of the two differently segmented versions of the German Oral History test set. The fine-segmented version is the default version used for all main experiments in the presented research work. The coarsely segmented data set is a preliminary version used primarily for initial experiments.

	Fine-Segmented	**Coarsely Segmented**
Length [hour]	3.52	3.53
Num. of Segments	2,392	102
Segment Length [s]	5.3 ± 3.5	124.5 ± 46.5
Words per Segment	11.6 ± 10.0	265.2 ± 91.2
Words per Second	2.1 ± 0.8	2.2 ± 0.5

To increase training data size and variance, we apply a method used in many recent Kaldi training routines by default: speed perturbation investigated by Ko et al. [2015]. The proposed approach by Ko et al. is to increase the data three-fold by creating two additional versions of each signal using the constant speed factors 0.9 and 1.1. We use this approach for all of our current and following experiments.

Evaluation Data

The evaluation of all models is performed on several evaluation sets from three domains to assure a reasonable estimate of the performance for real-world applications and to avoid selecting models that suffer from domain overfitting, cf. the discussion in Section 3.4. The evaluation sets used for these experiments are:

- the four *DiSCo* evaluation subsets, and the German Broadcast 2016 set for the German broadcast domain, cf. Section 3.4.3 and 3.4.4,
- the proposed German Oral History test set, cf. Section 3.4.6,
- the Interaction test set for the linguistics domain, cf. Section 3.4.7,

The results on German Broadcast 2016 and the Interaction test set were conducted after publication of [Gref et al., 2018a] and [Gref et al., 2018b] since these evaluation sets were not available early on—as described before.

To further study the influence of the segmentation length of the evaluation data on the different models, we compare results on two differently segmented versions of the German Oral History test set, as shown in Table 4.3. The coarsely segmented version of the test set has an average segment length of more than two minutes, making it nearly 25 times coarser-segmented than the fine-segmented version.

As described in Section 3.5, we use our default 500.000 words broadcast decoding language model for our experiments. We use a fixed subset of the GerTV1000h

corpus as a development set to adjust the language model weight to a fixed value during decoding for all test sets in each acoustic model experiment.

4.3.2 Results and Discussion

In the following, we present and discuss the results of the Kaldi acoustic model comparison experiments. We first report the results on the 128 hours training subset. Next, we analyze how the different models generalize when trained with 1000 hours of training data.

Acoustic Models Comparison with 128 Hour Training Data

The word error rates of the studied models on the several test sets are summarized in Table 4.4. The LF-MMI–TDNN-LSTM model using the training routine with per-frame dropout achieves the best results on all evaluation sets, except for Interaction. On the Interaction test set, this model achieves the second-best result after the CE-LSTM. Without per-frame dropout, the LF-MMI model performs significantly worse on all domains. The LF-MMI without per-frame dropout is often even worse than the CE-LSTM model, which is much simpler in terms of the neural network architecture and the training routine. In the following experiment with 1000 hours of data, we evaluate whether this is a general issue or a result of the reduced amount of training data.

The best LF-MMI and the CE-LSTM model both achieve a word error rate of 44 % on the German Oral History test set. Thus, both models beat the previous Audio Mining baseline by eleven percentage points while using just one-eighth of the training data, cf. Section 3.4.13. For the broadcast domain, the LF-MMI model even outperforms the literature results of the CTC-RNN model reported in Table 3.6—again, using just one-eighth of the training data. Therefore, we can assume a good generalization ability of these two models—although the accuracy on the challenging, non-broadcast domains still leaves a lot of room for improvement.

The bidirectional LSTM model performs worse than the other models on most test sets. Moreover, the results are significantly worse than the unidirectional LSTM. This observation is not entirely consistent with the results of the previously cited literature on bidirectional LSTM in ASR, which report generally improved performance. In our experiments, this is probably due to the halved projection dimension of the LSTM layers that we had to choose to keep the training possible in a reasonable amount of time. Thus, the results of LSTM and BLSTM are not entirely comparable. Because of the poor performance and the slow training, we do not consider the BLSTM in the subsequent experiments.

Table 4.4: Word error rates in percent achieved by the studied acoustic models trained on the 128-hour subset of the GerTV1000h corpus. DiSCo Average is the arithmetic mean of the results on the four DiSCo subsets.

	Cross-Entropy		**LF-MMI TDNN-LSTM**	
Test Set	**LSTM**	**BLSTM**	**Standard**	**Dropout**
GerTV Dev Set	16.6	17.2	17.0	**15.5**
DiSCo Average	17.2	18.2	17.9	**15.8**
Planned Clean	12.1	12.8	11.6	**10.6**
Planned Mix	16.7	17.6	17.4	**15.1**
Spontaneous Clean	13.9	14.6	13.8	**12.4**
Spontaneous Mix	26.0	27.6	28.9	**25.0**
German Broadcast 2016	15.1	14.9	14.6	**13.9**
Oral History	44.1	46.7	49.6	**44.0**
Interaction	**78.6**	82.0	83.0	79.7

Acoustic Models Comparison with 1000 Hour Training Data

The results for the models with 1000 hours of training data are shown in direct comparison to the results on 128 hours of training data in Table 4.5. Relatively, the *standard* LF-MMI–TDNN-LSTM model (without dropout) benefits most from the increased amount of training data, obtaining the highest relative and absolute decrease of the word error rates. The model outperforms the CE-LSTM model on most test sets when trained on 1000 hours of training data. However, the LF-MMI model with per-frame dropout still achieves the best results on all test sets. Thus, per-frame dropout helps to improve generalization when a smaller amount of training data is used—but it also generally improves performance when larger amounts of data are used.

Overall, the best LF-MMI model shows significant improvement for all domains and data sets by training on 1000 hours of broadcast speech. The model achieves a 13 % word error rate on average on the DiSCo subsets. For the easiest DiSCo subset *planned clean* the model achieves a word error rate below 10 %. For the most challenging DiSCo subset *spontaneous mix* it achieves a word error rate of 20 %. The remaining DiSCo subsets and broadcast test sets range from 11 % to 14 % word error rate. For oral history, a word error rate of just over 34 % is achieved, improving by ten percentage points—solely by adding more training data from

Table 4.5: Word error rates in percent achieved by the studied acoustic models increasing the training data from 128 hours to (↦) 1000 hours. For each test set, two word error rates are reported: the previous results of the models trained on a 128-hour subset from Table 4.4 on the left side of the arrow (↦) and those trained on the entire GerTV1000h corpus on the respective right side.

		LF-MMI	
Test Set	**CE-LSTM**	**Standard**	**Dropout**
GerTV Dev Set	16.6 ↦ 15.2	17.0 ↦ 14.5	**15.5** ↦ **14.0**
DiSCo Average	17.2 ↦ 14.7	17.9 ↦ 13.4	**15.8** ↦ **13.0**
Planned Clean	12.1 ↦ 10.6	11.6 ↦ **9.2**	**10.6** ↦ **9.2**
Planned Mix	16.7 ↦ 13.7	17.4 ↦ 12.9	**15.1** ↦ **12.0**
Spontaneous Clean	13.9 ↦ 12.1	13.8 ↦ **10.8**	**12.4** ↦ **10.8**
Spontaneous Mix	26.0 ↦ 22.3	28.9 ↦ 20.8	**25.0** ↦ **20.1**
German Broadcast 2016	15.1 ↦ 13.1	14.6 ↦ 12.5	**13.9** ↦ **12.2**
Oral History	44.1 ↦ 36.6	49.6 ↦ 36.6	**44.0** ↦ **34.3**
Interaction	**78.6** ↦ 71.1	83.0 ↦ 69.2	79.7 ↦ **67.6**

the broadcast domain. The improvement on Interaction is also significant, but performance on this data remains unsatisfactory.

Influence of Segmentation

To investigate the influence of segmentation on the different models, we evaluate the previously trained models with 128 and 1000 hours of training data with the coarsely segmented version of the German Oral History test set. We compare these results with the standard, fine-segmented version of the test set in Table 4.6. It is striking that the coarser segmentation only slightly degrades the cross-entropy trained models. This is especially true for the 128 hours trained CE-LSTM. The influence of the segmentation tends to become stronger for the 1000-hours-trained model.

The sequence-discriminative-trained LF-MMI–TDNN-LSTM models suffer significantly stronger from the long segments of the coarsely segmented test set than the cross-entropy models. Ultimately, the cross-entropy models achieve better results than the LF-MMI models on this version of the test set. The LF-MMI model trained with per-frame dropout is more affected by the long segments than the standard-trained LF-MMI model. In this case, per-frame dropout seems to be more harmful than beneficial. The effect is strongest for the LF-MMI model with dropout trained on 1000 hours, which previously achieved the best results on

Table 4.6: Comparison of coarse and fine segmented speech on the speech recognition performance using the oral history test set.

		Word Error Rate [%]		
	Segmentation:	**Coarse**	**Fine**	**Δ**
128 h	CE-LSTM	**45.3**	44.1	1.2
128 h	CE-BLST	48.2	46.7	1.5
128 h	LF-MMI–TDNN-LSTM	53.4	49.6	3.8
128 h	+ dropout	50.2	**44.0**	6.2
1000 h	CE-LSTM	**39.2**	36.6	2.6
1000 h	LF-MMI–TDNN-LSTM	43.3	36.6	6.7
1000 h	+ dropout	44.5	**34.2**	10.3

the fine-segmented Oral History test sets. The recognition performance decreases with longer segments by more than ten percentage points in absolute and by 30 % relative, respectively.

However, this limitation of the LF-MMI model with dropout is only of little detriment for application in a real-world system. For the application in systems such as Audio Mining, we have to apply an *utterance segmentation* that limits the segment length to a few seconds. The transcript can be concatenated after speech recognition to the desired segment length. The utterance segmentation can be realized by detecting speech pauses or spectral changes in the speech signal—similar to what is already realized for the speaker-change-aware segmentation in Audio Mining, cf. Section A.2.

However, for evaluation in Kaldi, where no automatic segmentation is performed, this peculiarity of LF-MMI Models should be taken into account, and suitable segment lengths should be chosen in advance to match the subsequent application. In the first publication of the present work, [Gref et al., 2018a], we used the coarsely segmented version of the Oral History test set for evaluation. Thus, the results there indicate that the CE-LSTM model is the model to be chosen for the oral history use case. In [Gref et al., 2018b], and all subsequent publications, we performed the analyses with the fine-segmented version that better reflects results in real-world applications.

4.3.3 Summary and Conclusion

The experiments presented in this study evaluated and compared different hybrid DNN-HMM acoustic models: a cross-entropy-trained LSTM and BLSTM hybrid

model and an LF-MMI–TDNN-LSTM model. The LF-MMI model was trained with and without per-frame-dropout. We trained all models with a 128 hours subset and the 1000 hours of German Broadcast training data to study how well the models generalize with different amounts of training data. We evaluated all models on a multitude of test sets from three different domains to get an appropriate estimate of real-world application performance and avoid the selection of models that suffer from domain overfitting. We have focused in particular on the German Oral History test set proposed in the presented research work.

In conclusion, the results indicate that the *standard* LF-MMI–TDNN-LSTM model does not generalize well when trained on only 128 hours of training data. A simple CE-LSTM with three layers achieves similarly good or better generalization and results on the different domains with the limited training data. The standard LF-MMI requires 1000 hours of annotated speech for training for good generalization. However, per-frame dropout helps elevate this problem when a smaller amount of training data is used—but it also generally improves performance when larger amounts of data are used. The LF-MMI–TDNN-LSTM model with per-frame dropout outperforms all other models in the 128 hours and 1000 hours training data scenario.

Furthermore, we studied the influence of the segmentation length of test data on the different models. We observed that very long segments degrade recognition performance for all models. However, the effect is relativity small for the cross-entropy models and stronger for the LF-MMI models. The LF-MMI model with per-frame dropout suffers the most from long segments, and the recognition performance deteriorates by up to 30 % relative. However, this is not critical for Audio Mining applications as long as utterance segmentation is used.

The LF-MMI model with per-frame dropout achieves a recognition performance of 34.2 % word error rate for oral history. The model also achieves the best recognition performance on all test sets from the other domains, beating previous literature benchmarks and indicating a decent real-world performance. Therefore, we use this model for all following experiments in the presented research work. For the subsequent experiments, we retrain this model with improved bootstrap training that further sightly improves the accuracy.

4.4 Study: Multi-Condition Training via Data Augmentation

After selecting a suitable acoustic model for the German oral history task, this section focuses on improving the acoustic robustness. In the systematic review of related works on challenges of oral history interviews for speech recognition in

Section 3.3.2, we discussed a multitude of distortions and challenges. We categorized the challenges into four categories and found acoustics to be one of the main challenges reported in the literature and our studied interviews. Additionally, we formulated the hypothesis that reverberation is one of the main acoustic challenges of our oral history interview collection. Thus, in this section, we discuss the influence of the acoustics of oral history interview recordings, especially of reverberation, on speech recognition in detail.

As described in Section 4.2, multi-condition training seems to be the most promising approach in the literature for single-channel recordings. Our study does not have appropriate in-domain training data to perform multi-condition training with real-world data and conditions. Our training data from the broadcast domain only covers certain conditions. Thus, there is a mismatch between our training data and the studied interviews. To overcome this limitation, we propose a combination of noise and reverberation data augmentation for multi-condition training to tackle this problem for German oral history interviews.

4.4.1 Recording Conditions of Oral History Interviews

Figure 4.3 visualizes the recording conditions of a typical oral history interview. Usually, a recording device is placed on a table in a small or medium-sized room, like a living room or kitchen. The interviewee is located at some distance to the recording microphone causing speech waves to be reflected on the walls and other surfaces in the room.

Reflections arriving at the microphone within 50 ms after the first wavefront are perceived as a single auditory event, cf. Avan et al. [2015]. These are called *early reflections* in contrast to *late reflections* that arrive after 50 to 80 ms. Early reflections cause *coloration* of the recorded sound that is perceived as a change of the timbre and is usually suppressed by binaural hearing, cf. Tsilfidis et al. [2013, pp. 359–364]. Thus, this effect becomes particularly evident in single-microphone recordings, such as in our interviews.

Idealized, the reflections in a room from a sound source (e.g., a speaker) to the receiver (e.g., a microphone or another person's ear) can be assumed to be a linear, time-invariant system. Mathematically, a complete description of this system is given by a *room impulse response* (*RIR*), cf. Tsilfidis et al. [2013, pp. 359]. The reverberation of a given speech signal can be modeled as a convolution of the speech signal s with the room impulse response h.

In some interviews, a noise source is also present in the recording. These noises can come from the street through a window, a noise source inside the room, or electronic devices. In Figure 4.3, this is symbolized by a loudspeaker. Depending

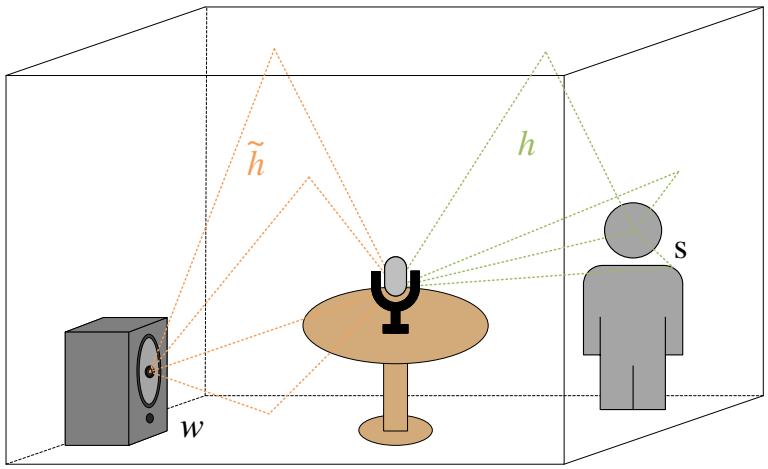


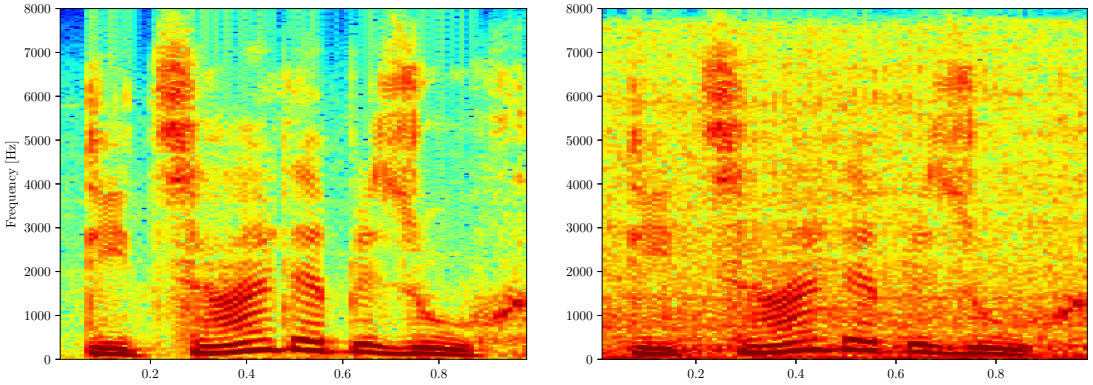
Figure 4.3: Strongly simplified visualization of the influence of room characteristics and microphone position on the recording by multiple, direct reflections of the speech sound waves during an interview. The loudspeaker symbolizes a noise source present during the recording, positioned at a location in the room different from the person speaking.

on the position in the room, the noise is also affected by the room reverberation. Since the position is often different from the interviewee, the effect of the room on the noise signal is modeled by another room impulse response $\tilde{h}$.

As described in Section 2.1.3, for GMM-HMM-based and hybrid acoustic models, usually MFCCs or related features are used. These features are based on the short-time Fourier-transformation spectrogram. Noises and room reverberation have a strong influence on the spectrogram. In Figure 4.4, we visualize the impact of reverberation, noises, and the combination of both, on the spectrogram.

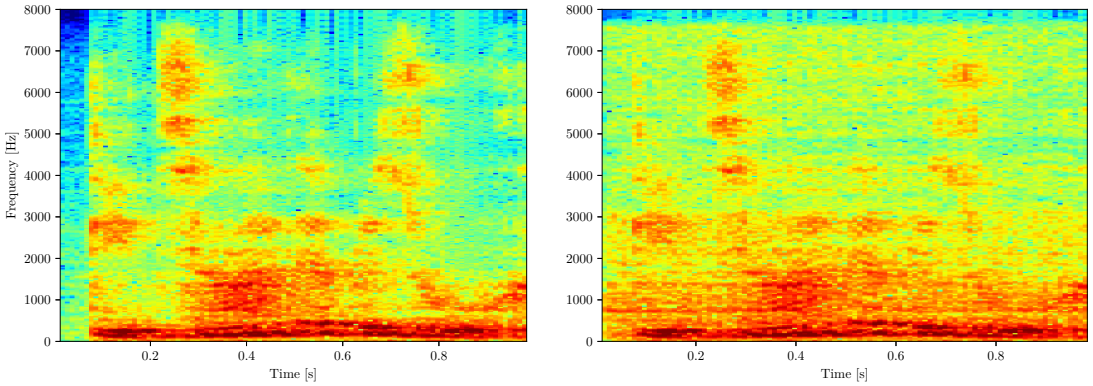
In spectrograms, harmonics caused by voiced speech sounds appear as equally spaced, relatively horizontal lines. Voiceless sounds appear as band-filtered noise. With good recording quality and intelligibility, these structures can be easily identified in the spectrogram—even at higher frequencies. This is well visible in Figure 4.4a, where the magnitude spectrogram of a *clean* speech signal from the broadcast domain is presented.

We use the data augmentation presented in the following Section 4.4.2 to visualize the influence of reverberation and additive noise. In Figure 4.4b, the influence of non-stationary, additive noise on the spectrogram is presented using a street noise signal from CHiME3 challenge data [Barker et al., 2015]. The noise overlays the speech signal over large frequencies ranges. This makes the identification of the speech components significantly more challenging. However, speech components with relatively high local signal energy can still be identified, such as the harmonics.



(a) Original, clean signal with good recording quality from the broadcast domain. The harmonics and the noise components of speech are both clearly visible, even for higher frequencies.

(b) Non-stationary street noise added to the clean signal with a 10 dB signal-to-noise ratio. Distortions are present over wide frequency ranges. Nevertheless, many harmonics remain visible due to the high local energy of voiced speech, e.g., in the time interval from 0.3 to 0.6 seconds. The harmonic structures are visible up to 3000 Hz.



(c) Artificial reverberation of an office room applied to the clean signal. The room reverberation acts like a low-pass filter, filtering harmonics and noise sounds and smearing the harmonics along the frequency axis at lower frequencies. Furthermore, the reverberation leads to a slight shift of the speech signal on the time axis and smearing along the time axis due to late reverberation, e.g., visible at 0.6 seconds.

(d) Application of reverberation from 4.4c after superposition with the noise from 4.4b. The low-pass effect of reverberation affects the noise signal in the same way as the speech signal. The noise components remain dominant up to about 1.5 kHz and can be seen attenuated up to about 3 kHz. As a result, the harmonics of speech become even more challenging to identify.

Figure 4.4: Magnitude spectrograms of a broadcast recording augmented in different ways to illustrate the effects of reverberation and additive noise on the spectrum. Additive noise, but especially room reverb, can blur harmonics and other speech characteristics and make them difficult to extract. All spectrograms use the same color bar for better comparison.

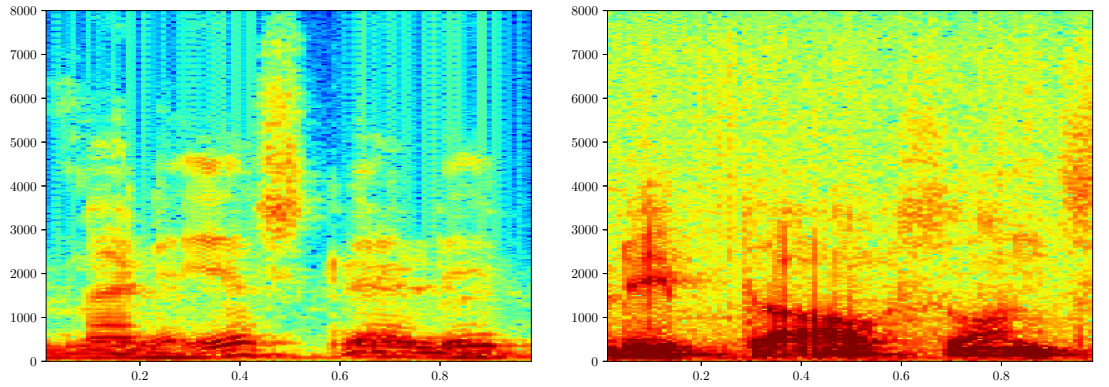
The influence of reverberation on the spectrogram is visualized in Figure 4.4c using a room impulse response of an office room from the *Aachen Impulse Response* (*AIR*) database [Jeub et al., 2009]. According to Jeub et al., the office room has an area of 32 square meters (5.0 m × 6.4 m) and a ceiling height of 2.9 meters. Therefore, it is more or less the size of a typical living room. The room impulse response used for Figure 4.4c was recorded at a distance of 2 meters from the source to the receiver. We consider this a realistic recording condition of a typical oral history interview. The room reverberation acts like a low-pass filter. It filters speech at higher frequencies and smears the harmonics along the frequency axis at lower frequencies, making them often indistinguishable.

In Figure 4.4d, the influence of both reverberation and noise on the speech spectrogram is presented. For this example, the room impulse response from Figure 4.4c is applied to the noisy speech signal from Figure 4.4b. The low-pass effect of the reverberation affects the noise spectrogram in the same way as the speech. Thus effect becomes evident considering the beginning of the signal between 0 and 0.05 seconds, where no speech is present. Up to about 1.5 kHz, the noise remains quite dominant, making it hard to distinguish speech from noise in the spectrogram. Like the speech signal, the local noise energy decreases with higher frequencies. It can be seen slightly attenuated up to about 3 kHz. Above 3 kHz, both the speech and the noise are strongly attenuated.

Figure 4.5 presents spectrograms of six different oral history interviews from our test set. Thus, these spectrograms show the real-world acoustic conditions of the interviews. Comparing these spectrograms with the clean speech recording from the broadcast domain in Figure 4.4a, the various auditory artifacts of the oral history interviews become evident. In almost all spectrograms, the influence of room reverberation is visible due to the before described low-pass filter effect. The interview in Figure 4.5a seems to be similar to the artificially applied reverberation from the broadcast example in Figure 4.4c. This is true both in terms of the spectrogram and subjective auditory perception. For the interview in Figure 4.5c, the low-pass filter effect is most evident. Above 2 kHz, the entire signal is almost completely suppressed.

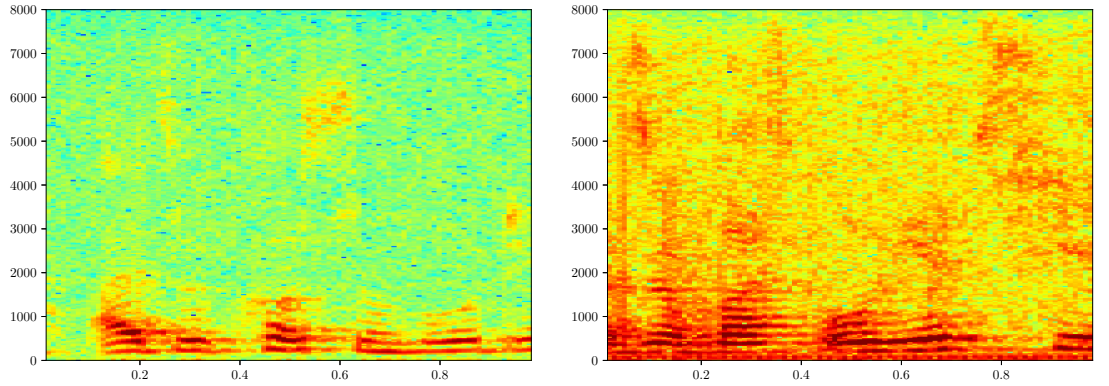
In addition to a slight room reverberation, stationary and non-stationary additive noises are visible in the spectrogram of the interviews in Figure 4.5d and Figure 4.5f. The interview from Figure 4.5d is the most similar to the artificial noisy broadcast example in Figure 4.4b. However, Figure 4.5f is more similar to the artificial combination of noise and reverberation in Figure 4.4d.

The interviews in Figure 4.5b and Figure 4.5e show other effects and artifacts visible in addition to these two effects. It is difficult to give a reason for these effects in retrospect since the recordings were made years or decades ago. Our hypotheses for these often non-linear, sometimes artificial sounding disturbances



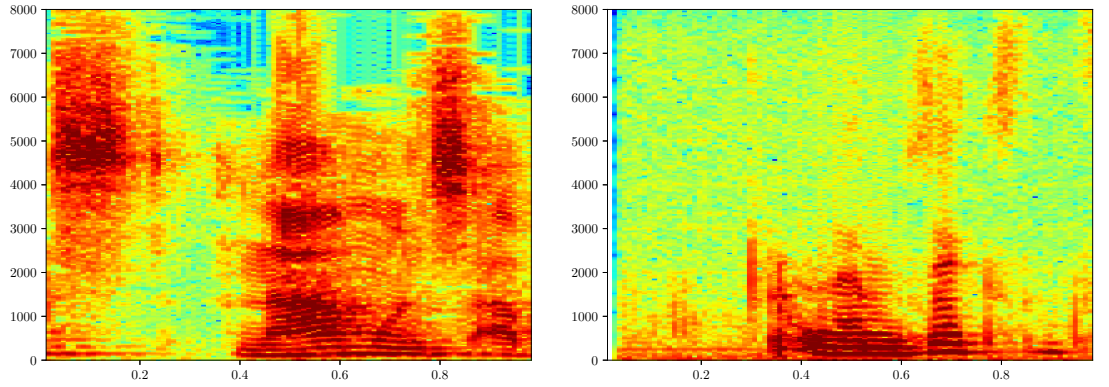
(a) No noise and some room reverberation, making the speech sound slightly dull. The reverberation damps harmonic parts of speech in higher frequencies.

(b) Moderate stationary noise and other artifacts—possibly due to aging of the magnetic tape and a low-quality microphone. Harmonics are barely discernible in the spectrogram.



(c) Low-energy stationary noise and room reverberation in what is probably a medium-sized room. Speech above 2 kHz is almost entirely suppressed.

(d) Stationary and electric humming noises, ticking clock in the background, slight room reverberation.



(e) Voice recorded very loudly. At times slight clipping occurs. No background noise is audible.

(f) Probably low-quality microphone, non-stationary background noise (probably street noises from an open window), and partial clipping.

Figure 4.5: Spectrograms of six different oral history interviews from the test set representing various acoustic distortions that occur to varying degrees. All spectrograms use the same color bar range for better comparison.

are the usage of low-quality microphones, clipping, and aging of the magnetic recording tape before digitization.

In conclusion, our qualitative comparison of the spectrograms of the artificially augmented broadcast recording with real-world oral history interviews confirms the hypothesis made in advance that a multitude of different acoustic distortions co-occurs in oral history interviews with varying degrees. Reverberation is one of the distortions that seem to occur in most interviews to some degree. The reverberation of the small and medium-sized rooms in which the interviews were recorded manifests itself by a low-pass filter effect and smearing of harmonics along the frequency axis in the spectrogram.

Furthermore, stationary and non-stationary additive noises appear in many interviews to some degree. Reverberation often influences these noises similar to speech. Data augmentation seems to replicate these effects well and is a promising approach to improve robustness. We conclude that a combination of both effects is beneficial for data augmentation for the studied oral history interviews. However, other acoustic effects can only be simulated to a limited extent since the exact origin is unclear. These will be further addressed in Chapters 5 and 6 using transfer learning and model adaptation.

4.4.2 Noise and Reverberation Data Augmentation for Oral History Interviews

This section lays out our proposed data augmentation approach for multi-condition training to improve the robustness of hybrid acoustic models for German oral history interviews. Our proposals are primarily based on the systematic literature review from Section 3.3.2 and the qualitative analysis and comparison of German oral history interviews with data augmented broadcast spectrograms in Section 4.4.1. In summary, we propose:

- the combination of noise and reverberation data augmentation of the (clean) broadcast training data to obtain a wide range of different, realistic recording conditions for our interviews.
- using the augmented training data not only during neural network training but for the entire GMM-HMM bootstrap model training. Other approaches in the literature often use clean training data for the GMM-HMM bootstrap pipeline and apply data augmentation only in the neural network training state.
- randomly overlapping different noise signals to generate *unique* noise signals for each speech signal in the case that only an insufficient amount of noise data is available compared to the speech data.

- combining clean and augmented speech during training. This is based on observations in preliminary data augmentation experiments, where only augmented and no clean data was used. These experiments resulted in decreased recognition performance.

In general, these proposals aim to improve robustness and generalization with *multi-condition* training. In the following, we describe these proposed aspects in more detail.

Combination of Noise and Reverberation Data Augmentation

Based on the observations in the previous section, we propose combining noise and reverberation data augmentation for multi-condition training to improve robustness for LF-MMI acoustic models for German oral history interviews. The proposed data augmentation is visualized in Figure 4.6. Each speech signal is augmented by convolution with a randomly selected room impulse response. A room impulse response from the same room but at a different, random position is used to augment a randomly selected noise signal. After applying reverberation, the noise signal is added to the speech with a signal-to-noise ratio randomly drawn from a predefined range.

Mathematically, the proposed augmentation can be described as

$$(x_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}} + \lambda \cdot (w_n)_{n\in\mathbb{N}} * (\tilde{h}_n)_{n\in\mathbb{N}}, \tag{4.1}$$

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

As observed in the previous section, some interviews are distorted by reverberation, but no background noise seems to be present. If only reverberation and no background noise affects the speech signal, $\forall n \in \mathbb{N} : w_n = 0$ applies and yields

$$(x_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}} \tag{4.2}$$

for data augmentation.

Multi-condition training with data augmentation using reverberation and noises was already studied in the days of GMM-HMMs, for example, by Sehr et al. [2011]. However, with the emergence of hybrid DNN-HMM, it is again the object of investigation for neural-network-based acoustic models. For example, this becomes evident in the meta-analysis by Kinoshita et al. [2016] of the contributions for the REVERB challenge. Kinoshita et al. identify the following two aspects which in-

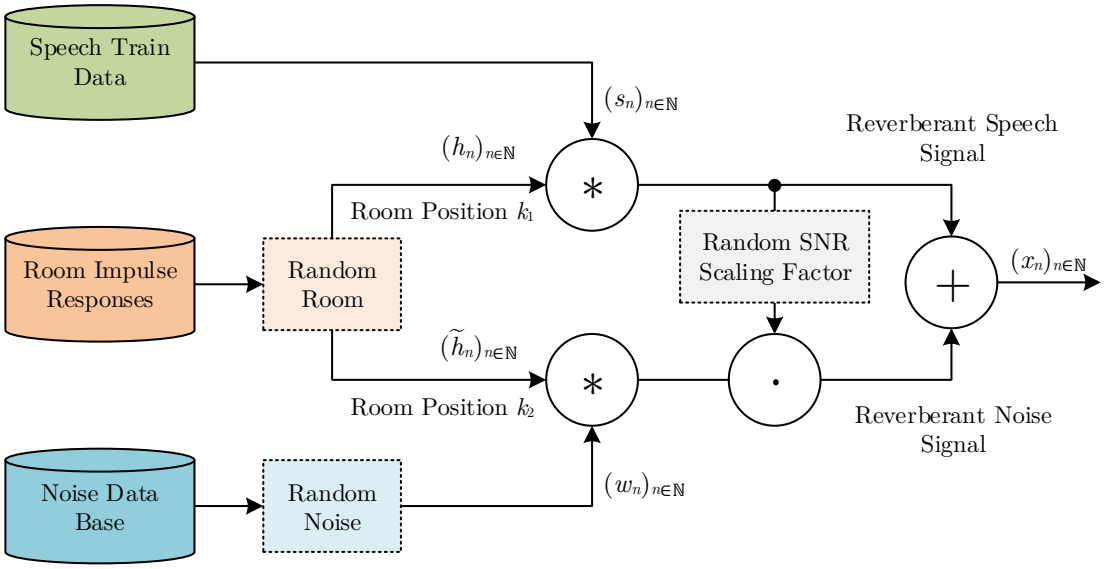


Figure 4.6: Visualization of the noise and reverberation data augmentation. Reverberation is applied to the speech and noise signal before both are added with a random signal-to-noise ratio. The room impulse responses are chosen randomly with the condition that they are from the same room, but they can be from different positions. This aims to match the actual recording conditions of oral history interviews.

fluence the overall performance in reverberant speech recognition the most: DNN-based (hybrid) acoustic models and multi-condition instead of clean training data.

As discussed earlier, a limitation in the literature for real-world applications is that only exactly one specific challenge or distortion is often examined. This is also true for the REVERB challenge, where reverberated speech is studied in almost noise-free conditions (no noise or with 20 dB signal-to-noise ratio) and on limited vocabulary, read speech task corpora [Robinson et al., 1995; Lincoln et al., 2005].

Different research works explore the combination of different challenges and data augmentation approaches to address this limitation in the literature at the beginning of the presented research work. For instance, Hartmann et al. [2016] studied the combination of noise data augmentation with speed perturbation and speaker-adaptive augmentation for low-resourced languages.

Also, multi-condition training with noise and reverberation data augmentation was applied and studied in real-world applications at that time. For instance, among multiple proposed approaches in the pipeline, Li et al. [2017] use artificial room impulse responses from a room simulator and real-world noises, including music and ambient noises, to train the Google Home system for far-field, multi-channel speech recognition.

A combination of noise and reverberation data augmentation, similar to the one we propose for our use case, is studied by Ko et al. [2017] roughly at the same time as our work. The authors use an equation similar to Equation 4.1 to model the influence of noises and reverberation but with a sum of noise sources from multiple positions. In our approach, we only consider one noise source position.

Ko et al. studied data augmentation with simulated and real-world room-impulse responses and different noise types for far-field ASR with cross-entropy and LF-MMI models. The approaches are evaluated on the English far-field large vocabulary continuous speech recognition *ASpIRE challenge* [Harper, 2015] and *AMI meeting corpus* [Mccowan et al., 2005; Hain et al., 2006]. The data augmentation was applied on recordings from the English Switchboard [Godfrey et al., 1992] and Fisher corpus [Cieri et al., 2004]. Additionally, the authors evaluated their system on the close-talk telephone speech from Switchboard. The authors concluded that real room-impulse responses perform slightly better than simulated impulse responses. To overcome this, the authors propose using noises with simulated room-impulse responses.

Ko et al.'s results convincingly show how reverberation data augmentation can help to adapt the training data for far-field speech recognition tasks with hybrid acoustic models—and in addition, it can also improve the performance of speech recognition in general. However, from our task's perspective, a limitation of the study is that the studied challenges cover only room reverberation but no noises and other acoustics distortions we observe in our interviews.

Multi-Condition Bootstrap Training

Ko et al. use a clean-trained bootstrap pipeline and apply the multi-condition data only for neural network training. Thus, the phonetic decision tree generation, context-dependent state-tying for the GMMs, and other aspects of bootstrapping are performed on clean data. The alignment of the data is performed with a clean-trained GMM-HMM system. While this saves a lot of computational time, this might be a limitation since the GMM-HMM systems used might not be as robust against acoustic distortions and lead to poorer alignments.

For distortions without temporal influence on the speech signal, such as additive noises, it is conceivable to perform the alignment only for the clean data and use these alignments for the corresponding augmented speech samples. However, reverberation leads to shifts on the time axis and smearing of the signal along time due to late reverberation. Both effects are also observable in small and medium-sized rooms, e.g., as shown by comparing Figures 4.4a and 4.4c. Therefore, we propose and study the usage of augmented data in bootstrap training to obtain robust GMM-HMM acoustic models to obtain more precise alignment for distorted data. We expect that this ultimately leads to improved speech recognition per-

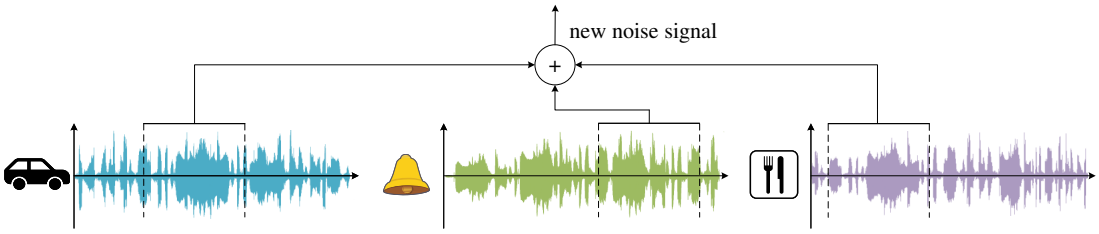


Figure 4.7: Graphical representation of the procedure for generating a new, previously unused noise signal for each speech sample. Random time points are selected from three different noise recordings from the database to create new noise combinations repeatedly.

formance of the hybrid acoustic model trained on top of these GMM-HMMs for speech recorded in challenging acoustic conditions.

Random Noise Overlapping

We had comparatively little noise data available for the first experiments with data augmentation with real-world noise recordings. For the 1000 hours of speech data, only 14.5 hours of noise were available. If the noise were applied to the entire data, each noise would be seen in training about 70 times per epoch. Overfitting to this noise would be expected. Therefore, we used a method to artificially create a new noise signal from the existing noises for each speech sample that has not been used in exactly this way before.

The procedure is illustrated in Figure 4.7. We randomly select three noises from the database for each speech signal. A random time point in each signal is selected from which a portion of the signal is extracted. Another noise signal is appended if the extracted signal is shorter than the current speech signal. These three randomly selected noise signals are added to create a new noise signal.

The resulting noises often have a stationary rather than a non-stationary character. This is due to the averaging effect of a superposition with three different noises. Training with this data, we expect the robustness to increase, especially for oral history interviews. We perform comparative experiments with simple stationary additive white Gaussian noise (AWGN) to evaluate the influence of real-world noises created with our approach.

Combining Clean and Augmented Speech During Training

We observed in preliminary experiments on data augmentation with CE-LSTM models and the 128-hour training data set that training hybrid acoustic models only on reverberated or noisy speech generally tends to lead to decreased instead

of increased recognition accuracy. This observation tends to be true for both broadcast recordings and oral history interviews.

This behavior is probably because this type of training defects the goal of *multi-condition* training in ASR and is rather *matched-condition* training. Matched-condition training aims to train systems on data that *matches* the evaluation or inference data. However, an exact replication of the distribution of acoustic distortions in oral history interviews is not feasible. As explained before, the frequency and type of distortions are different in each interview—and the recording situation cannot be reconstructed in retrospect. Thus, in multi-condition training with data augmentation, the aim is to obtain training data that do not precisely *match* but *cover* the evaluation or inference conditions, cf. Vincent et al. [2017].

Thus, we propose combining noise and reverberation data augmented speech with original *clean* data for training to increase robustness and cover an extensive range of acoustic conditions. This is in line, e.g., with Ko et al. [2017], who state that combining clean and reverberated training data also leads to considerable improvements for close-talking scenarios in their experiments.

Additionally, as for the first experiments in Section 4.3, we apply speed perturbation techniques to increase training data variance by creating two additional versions of each signal using the constant speed factors 0.9 and 1.1.

4.4.3 Experimental Setup

In the following, we lay out our experimental setup to evaluate the noise and reverberation data augmentation. Again, we conduct all experiments with the Kaldi ASR toolkit [Povey et al., 2011].

DNN-HMM Acoustic Models

For evaluation, we use the best-performing approach from the acoustic model comparison experiments in Section 4.3: the LF-MMI–TDNN-LSTM acoustic model trained with per-frame dropout. In an ablation study laid out later, we perform an additional comparison experiment with the CE-LSTM model to demonstrate that the data augmentation improves the robustness for oral history interviews in a model and training-criterion independent manner.

Bootstrap Training

In Section 4.3, we compared models with varying training data sizes. In the current and the subsequent experiments, we train the LF-MMI model with the entire 1000 hours of training data—or more. We increase the size of the subsets during bootstrap training for the LF-MMI model, as shown in Table 4.7, to account

Table 4.7: Training data subsets used for bootstrap training in the previous and the current setup. The training data size for the first three GMM-HMM models is roughly tripled. Feature and tied state configuration are the same as in the previous setup, cf. Table 4.2.

Model	Previous Setup	New Setup
Monophone	11,000 shortest utt.	30,000 shortest utt.
Triphone 1	38,000 rand. utt.	100,000 rand. utt.
Triphone 2	38,000 rand. utt.	100,000 rand. utt.
Triphone 3	entire data	entire data
Triphone 4	entire data	entire data

for this increased training data. The features and tied-state configurations are not changed. This new setup is more closely related to the Kaldi setup for English Switchboard models. We expect this setup to improve the alignment (used for chunking in LF-MMI training) and state-tying, which should ultimately improve the recognition performance of the hybrid model.

Training Data

We used 266 room impulse responses of real small and medium-sized rooms for the reverberation data augmentation. This database is obtained by combining a subset from the *Aachen Impulse Response* (*AIR*) database [Jeub et al., 2009] with in-house room impulse responses.

In our experiments, we study two noise types: artificially created additive white Gaussian noise and 14.5 hours of real-life noise recordings. For the real-world noise experiments, we combine in-house data with 8 hours of data from the CHiME3 challenge, recorded in noisy environments (on a bus, in a cafe, pedestrian area, and street junction) [Barker et al., 2015]. The in-house data contains recordings of different noise sources, such as sirens, hairdryers, crowd cheering, and noises from kitchen devices.

We augment the entire GerTV1000h training data set with the room impulse response and noise databases. In the first step, we create three data augmented versions of the corpus:

- **Reverb+RealNoise**: All signals are augmented according to Equation 4.1. We apply the random noise overlapping to real-world noise recordings to generate new noise signals for each speech recording. As described in Section 4.4.2, noise signals are also reverberated using a room impulse response from the same room but a different location. After reverberation, we apply a

Table 4.8: Overview of the different multi-condition training setups used for noise and reverberation data augmentation experiments. The amount of training data is kept fixed to the original amount of data for all Mix models. As the name implies, for 3-fold, the amount of training data is increased 3-fold.

	Original	**Reverb**	**Reverb+ WhiteNoise**	**Reverb+ RealNoise**
Clean	100 %	0 %	0 %	0 %
Mix-Reverb	50 %	50 %	0 %	0 %
Mix-Reverb+WhiteNoise	40 %	40 %	20 %	0 %
Mix-Reverb+BothNoises	35 %	35 %	15 %	15 %
Mix-Reverb+RealNoise	40 %	40 %	0 %	20 %
3-fold*	100 %	100 %	0 %	100 %

random signal-to-noise ratio between 10 and 20 dB. This range roughly corresponds to the perceived noise range of typical interviews, from only slightly perceptible to moderate noise.

- **Reverb+WhiteNoise**: Similar to Reverb+RealNoise but instead of real-world noises, additive white Gaussian noise is used with the same setup.

- **Reverb**: For this augmentation, no noise and only reverberation is used according to Equation 4.2.

Based on these augmented versions of the GerTV1000h corpus, we created different multi-condition training sets randomly selecting samples from the corpora versions using different distributions, as shown in Table 4.8. We study four different mixtures of clean and augmented speech recordings where the overall amount of training data is kept fixed to the original setup. Mix-Reverb focuses on reverberation only, combining clean with reverberated speech to equal extents. This mixture does not contain any additional noises and serves as a comparative experiment for the influence of noise on multi-condition training. Mix-Reverb-+WhiteNoise adds AWGN to the setup, while Mix-Reverb+RealNoise adds real-world noises. Mix-Reverb+WhiteNoise serves as a comparison experiment for the real-world noise setup Mix-Reverb+RealNoise. In Mix-Reverb+BothNoises, the share of noises is sightly increased compared to Mix-Reverb+WhiteNoise and Mix-Reverb+RealNoise, combining AWGN and real-world noises.

Additionally, in later experiments that finished after publishing [Gref et al., 2018b], we performed experiments where the amount of training data is increased 3-fold using the proposed noise and reverberation data augmentation. The 3-fold

setup reassembles and extends the concept of Mix-Reverb+RealNoise. It combines the entire Clean, Reverb, and Reverb+RealNoise versions of GerTV1000h. The model was first introduced as one part of a larger pipeline in [Gref et al., 2019].

Evaluation Data

We use our default 500.000 words broadcast decoding language model for our experiments, cf. Section 3.5. Again, we use the fixed-sized subset of the clean GerTV1000h corpus as a development set to adjust the language model weight to a fixed value for each acoustic model during decoding all test sets.

The evaluation of the experiments is performed on the same evaluation sets as in the previous set of experiments in Section 4.3: the DiSCo subsets and German Broadcast 2016 for the broadcast domain, the Oral History test set, and the Interaction test set. These sets cover three different domains to assure a reasonable estimate of the performance for real-world applications and avoid domain overfitting. The main focus is the German Oral History test set since it represents the studied use case.

Additionally, we evaluate the models of the main experiments on the acoustic robustness with additional test sets: Challenging Broadcast (cf. Section 3.4.5) and Spoken QALD-7 (cf. Section 3.4.8). These sets cover challenging acoustic recording conditions for different domains. The results on these sets were conducted after the publication of [Gref et al., 2018b] as these evaluation sets were not available early on.

Ablation Studies

In Section 4.4.2, we proposed multi-condition bootstrap training instead of training only the neural network acoustic model on multi-condition data. To verify the improvement of the robustness of the final acoustic model, we perform an ablation study. We use the clean-trained bootstrap and apply the multi-condition training only in the neural network training stage.

Additionally, we perform an ablation where we replace the acoustic model with the CE–LSTM from Section 4.3, trained on the 128-hour subset of the multi-condition data—using the same segments of GerTV1000h used in Section 4.3. This aims to demonstrate that the improvement of the acoustic model's robustness is not depended on the selected model, the training criterion, and a large amount of training data, but is a property of the proposed data augmentation.

Table 4.9: Comparison of the previous and updated GMM-HMM bootstrap training on the final LF-MMI-TDNN-LSTM model (with per-frame dropout). Results are reported as word error rates in percent. Increasing the subsets in bootstrap training of GMM-HMM models in the new setup leads to consistent improvement of the hybrid model on all domains.

Test Set	Old	New
GerTV Dev Set	14.0	**13.8**
DiSCo Average	13.0	**12.8**
Planned Clean	9.2	**9.0**
Planned Mix	**12.0**	**12.0**
Spontaneous Clean	10.8	**10.6**
Spontaneous Mix	20.1	**19.7**
German Broadcast 2016	**12.2**	**12.2**
Oral History	34.3	**34.2**
Interaction	67.6	**65.7**

4.4.4 Results and Discussion

In the following, we present and discuss the results of the different multi-condition experiments. We first report the results of the new bootstrap training on clean data only. This model serves as a new baseline for the subsequent experiments, where we compare and discuss the different multi-condition setups for the LF-MMI model. In the two ablation studies, we first analyze the influence of multi-condition data during bootstrapping. Then, we demonstrate the robustness of the proposed data augmentation with a CE-LSTM model trained on fewer data.

New Clean-Trained Baseline with Updated Bootstrapping

Table 4.9 compares the results of the LF-MMI–TDNN-LSTM model (trained with per-frame dropout) for the previous and new bootstrap configuration. The new bootstrapping utilizes larger subsets of the GerTV1000h corpus for the GMM-HMM training, cf. Table 4.7. The results show a small yet consistent improvement in recognition performance for all domains.

It is likely that further increasing the amount of training data in bootstrapping can lead to further improvement in the overall recognition performance. However, this comes at the expense of the increased training time and computational load in trade of a presumably comparatively small improvement. Therefore, we refrain from this in the presented research work and use the new bootstrap configuration for all subsequent experiments.

Table 4.10: Results on different domains of the multi-condition trained LF-MMI–TDNN-LSTM models. Results are reported as word error rates in percent. DiSCo Average is the arithmetic mean of the results on the four DiSCo subsets. The best result with fixed training data size, i.e., without 3-fold, is highlighted for each test set. Additionally, it is highlighted when 3-fold achieves the overall best result.

Test Set	Clean	Mix-Reverb	Mix-Reverb +WhiteNoise	Mix-Reverb +BothNoises	Mix-Reverb +RealNoise	3-fold
GerTV Dev Set	13.8	13.9	13.9	**13.6**	13.9	13.7
DiSCo Average	12.8	12.5	12.4	12.5	**12.3**	**11.8**
German Broadcast 2016	12.2	12.2	12.1	**12.0**	12.3	**11.5**
Challenging Broadcast	21.2	20.7	20.4	20.5	**20.2**	**20.1**
Oral History	34.2	30.6	30.3	29.6	**29.5**	**28.2**
Interaction	65.7	53.1	50.8	49.9	**49.8**	**47.8**
Spoken QALD-7	20.9	**17.9**	19.2	19.0	18.6	18.3

Result Overview for Multi Condition Training with Noise and Reverberation Data Augmentation

In Table 4.10, we report the results of the LF-MMI–TDNN-LSTM models trained with the different multi-condition training data sets compared to the clean-trained baseline. We first discuss the multi-condition training setups of the Mix models with the same training data size as the clean-trained baseline. The 3-fold model, i.e., the extension of the Mix models with three-fold augmented training data, is discussed and presented later in this section after the Mix models.

Overall, multi-condition training with data augmentation with fixed training data size improves performance on all test sets, both in broadcast and the other three domains. However, the improvements in the broadcast domain are not as substantial as for the out-of-domain data. In the best cases, the word error rate on the broadcast test sets improves by 2 to 5 % relative to the clean-trained baseline. For our oral history data, we observe an improvement by 14 % relative to the baseline with Mix-Reverb+RealNoise. Spoken QALD-7 shows a similar relative improvement in the best case with Mix-Reverb, and 11 % relative improvement with Mix-Reverb+RealNoise. The best relative improvement for the Interaction test set is 24 % with Mix-Reverb+RealNoise. This indicates that the proposed

data augmentation can successfully compensate the domain mismatch between broadcast and the three other domains to a certain extent and at the same time improve the robustness in the original domain—without increasing the amount of training data. However, the remaining substantial gap in the word error rate between the broadcast domain and the other domains indicates that there continues to be a domain mismatch by the models.

Overall, combining clean data with real-world noise and room reverberation in Mix-Reverb+RealNoise (cf. Table 4.8) yields the best results among the multi-condition setups with fixed training data size for four out of six evaluation sets. Exceptions to this are German Broadcast 2016 and Spoken QALD-7. In the case of German Broadcast 2016, the overall impact of multi-condition training appears to be relatively small (0.5 % relative on average), so this might also be considered statistical noise. However, this is not true for Spoken QALD-7. Multi-condition training with noise and reverberation data augmentation significantly improves the overall recognition performance compared to the clean-trained baseline on this test set. However, Mix-Reverb achieves by far the best result. One possible reason is that Mix-Reverb best represents the recording conditions of Spoken QALD-7. This test set was recorded by various people with their respective laptop recording equipment. Noise is hardly encountered in the recordings. However, with microphones built into the laptop, a distance to the speaker is common, leading to a similar room reverberation as represented in Mix-Reverb. Among the multi-condition setups with noises, setup Mix-Reverb+RealNoise with real-world noise achieves the best results on Spoken QALD-7.

Furthermore, the results confirm an observation made in various of our preliminary experiments: The comparison of systems solely on the GerTV Dev Set does not lead to a reasonable estimation of the real-world performance of the systems. This set shows different trends than observed on the test sets. For three of the multi-condition setups, the results on the development set worsen instead of improving—as the results do for the test sets. This is not only true for the out-of-domain data but also for data from the broadcast domain. We assume this is because the development set is a split of the clean GerTV1000h training data and thus precisely represents these conditions. Evaluation with only this data set would thus be susceptible to domain overfitting, cf. the discussion in Section 3.4. It should be refrained from performing a model selection solely on this GerTV development set. Furthermore, the choice of the language model weight for decoding, which we have performed on this set so far, might not yield the best possible results for real-world data. Therefore, in the following chapters, we choose a fixed, standard language model weight for most experiments.

The 3-fold model finished training much later than the other systems due to the tripled training data. The model can be understood as an extension of Mix-Reverb-

Table 4.11: Word error rate (in percent) of the multi-condition trained LF-MMI–TDNN-LSTM models on the four DiSCo subsets. The best result with fixed training data size, i.e., without 3-fold, is highlighted for each test set. Additionally, it is highlighted when 3-fold achieves the overall best result.

Test Set	Clean	Mix-Reverb	Mix-Reverb +WhiteNoise	Mix-Reverb +BothNoises	Mix-Reverb +RealNoise	3-fold
DiSCo Average	12.8	12.5	12.4	12.5	**12.3**	**11.8**
Planned Clean	9.0	**8.8**	**8.8**	9.2	8.9	9.0
Planned Mix	12.0	11.4	**11.2**	**11.2**	11.3	**10.8**
Spontaneous Clean	10.6	**10.2**	10.7	10.5	10.4	**9.9**
Spontaneous Mix	19.7	19.5	19.0	19.0	**18.7**	**17.5**

+RealNoise, as it represents similar conditions—just with three-fold increased training data size. The model consistently outperforms Mix-Reverb+RealNoise on all test sets and achieves the best result on five of the six test sets. The relative improvement in the broadcast domain with 3-fold is more substantial, with 5–8 % relative improvement to the clean-trained baseline. Only on Spoken QALD-7, Mix-Reverb remains the best model. However, the 3-fold model further reduces the gap to Mix-Reverb on this set to 0.4 % absolute. With this model, we achieve a 28.2 % word error rate on our oral history target domain, i.e., slightly less than 18 % relative improvement to the clean-trained baseline.

Discussion of Multi Condition Training on Broadcast Recordings

Table 4.11 further disaggregates the results of the models on the DiSCo data for the four subsets. This analysis aims to evaluate the different effects of augmentation on clean and acoustically distorted broadcast recordings. Even though this is not the target domain of the presented research work, the analysis shall help better assess the real-world performance of the systems in different conditions.

On the clean subsets, the model trained on Mix-Reverb achieves the best results. Mix-Reverb contains no additional noises and only combines clean with reverberated speech. The addition of noises in the other multi-condition data sets degrades the recognition performance on these subsets compared to Mix-Reverb. However, the results on Mix-Reverb+RealNoise are only slightly worse than on Mix-Reverb and better than the multi-condition setups with stationary AWGN.

At the same time, adding noises to the multi-condition training substantially improves the robustness of the model on the DiSCo Mix subsets. These substantial improvements lead to the fact that the model training on Mix-Reverb+RealNoise achieves the best result on average for the DiSCo data even though it is only the best performing system on one of the four subsets.

The 3-fold training yields another substantial improvement in recognition performance. The model not only achieves better results on the Mix subsets. The results also indicate improvements by the model for spontaneous speech. To the best of our knowledge, this is the first time a word error rate below 10 % is achieved on DiSCo Spontaneous Clean without DNN-LM rescoring.

Discussion of Multi Condition Training on Oral History Interviews

The oral history interviews in our test set have a wide range of recording conditions and signal quality. Therefore, the overall word error rates only give an averaged indication of the improvement achieved by the different multi-condition setups. To assess the performance of the multi-condition setups for the different interviews, a box plot diagram of the word error rate of each of the 35 oral history interviewees in the test set is given in Figure 4.8.

The box plots reveal a wide range of word error rates for the different interviews. Without data augmentation, the distribution is positively-skewed towards lower error rates. Thus, the median is lower than the average word error rate. Half of the interviews have a word error rate of 18.4–31.7 %, the other half in the range 31.7–59.6 %.

In general, the word error rate decreases for all quartiles with multi-condition training, especially for high word error rates. For the clean model, more than a quarter of all interviews—11 to be exact—have a word error rate above 40 %. For the multi-condition models, only a few interviews are above this error rate: six interviews for Mix-Reverb, four interviews for the three Mix models with noises, and three interviews for 3-fold.

Utilizing real-world noises in training with seems further to improve the recognition for these very challenging interviews. With Mix-Reverb+BothNoises and Mix-Reverb+RealNoise, three-quarters of all interviews have an error rate below 35 %—and below 33 % for 3-fold—which we consider a substantial improvement compared to the baseline. With these three models, the error rate of all interviews drops below 45 %—except for one outlier that almost consistently has a word error rate of roughly 60 %. But even for this interview, the real-world noise setups best improve the recognition performance. The 3-fold model lowers the word error rate from 59.6 % to 52.8 %. This indicates that acoustic conditions are one reason for the poor recognition performance on this interview. However, other challenges in the recording still make recognition extremely difficult, such as

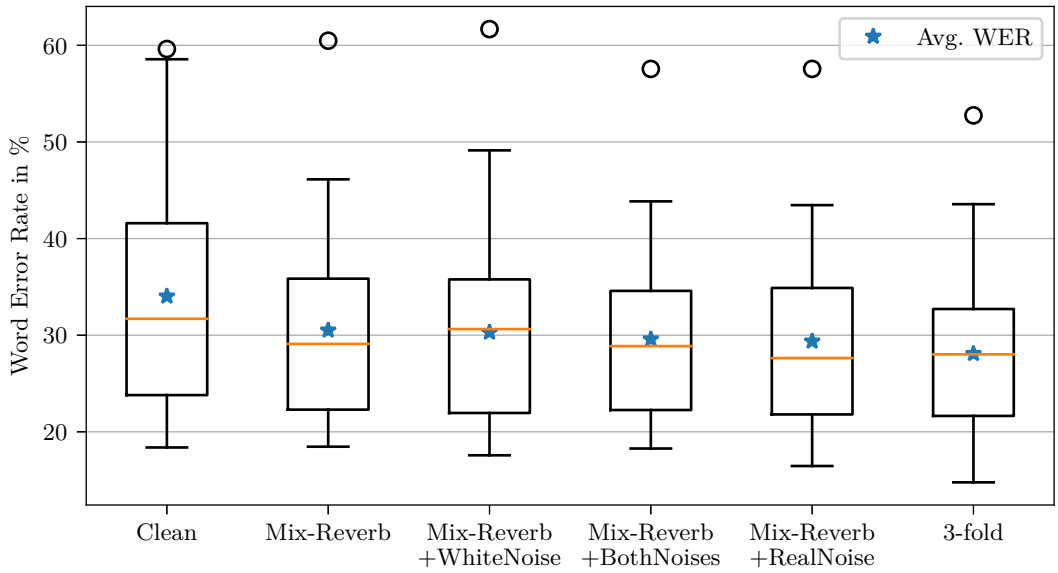


Figure 4.8: Box plot diagram of the word error rates of the 35 different oral history interviews for each model in the multi-condition experiments. The interquartile range is set to 1.0 so that the same outlier is plotted for all box plots. The whiskers represent the minimum to maximum word error rate, excluding the outlier. The lower and upper lines of the boxes represent the first and third quartiles of the word error rate. Half of the interviews have word error rates between these two lines. The inner lines of the boxes represent the median.

spontaneous speech, unclear pronunciations, a dialect, and other acoustic artifacts in the recording not modeled by the data augmentation. Particularly, the 3-fold model also improves the recognition performance for interviews with a low error rate.

The box plots only show the overall distribution of word error rates but not the improvement or deterioration for individual interviews. Even though most results improve, the recognition performance also deteriorates for some interviews due to multi-condition. In Mix-Reverb, five of the 35 interviews have a higher word error rate than the clean-trained baseline. Mix-Reverb+WhiteNoise is the least robust approach, with seven interviews recognized worse than the baseline. This could also be a reason for the slightly deviating median in Figure 4.8 for Mix-Reverb+WhiteNoise compared to the other models. There are only two interviews in the intersection of the deteriorating interviews in Mix-Reverb and Mix-Reverb+WhiteNoise. This indicates that the models generalized to different acoustic conditions. In Mix-Reverb+BothNoises, there are only three interviews with a worse error rate than the baseline. Mix-Reverb+RealNoise and 3-fold show the most robust improvements, with only two interviews getting worse than the baseline.

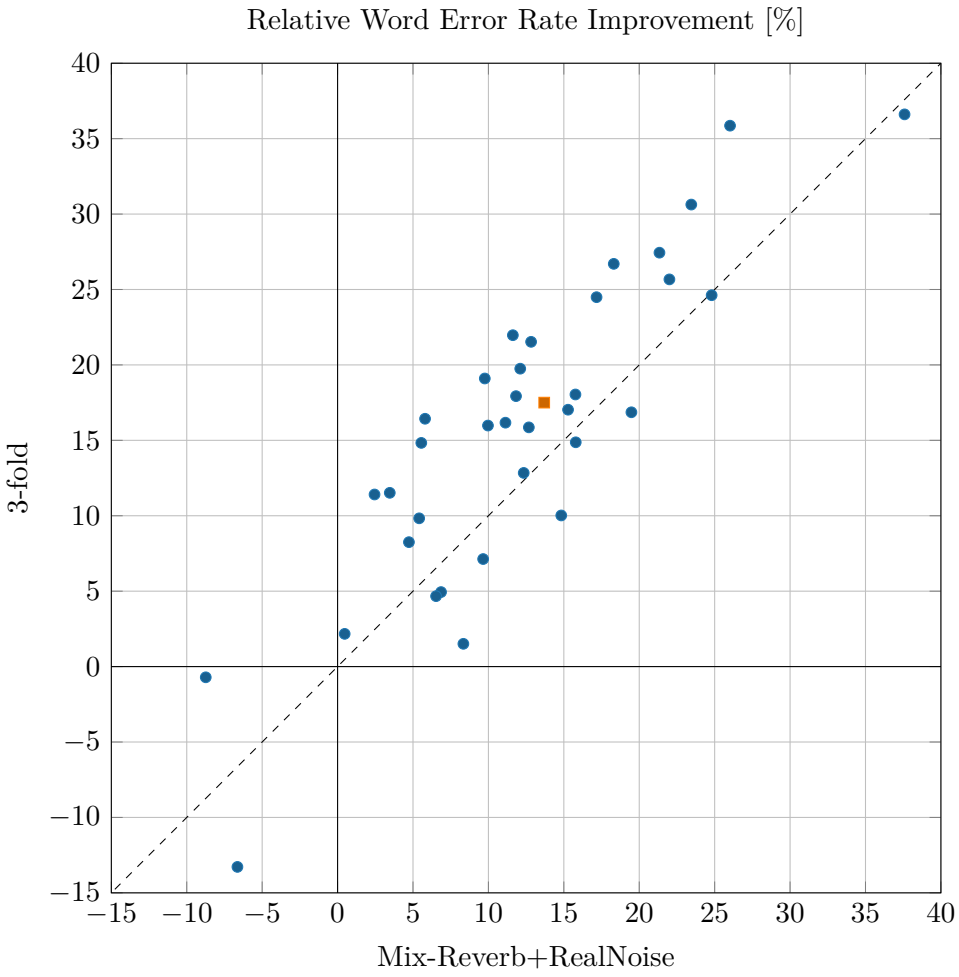


Figure 4.9: Scatter plot of the relative word error rate improvements compared to the clean-trained baseline of Mix-Reverb+RealNoise and 3-fold. All points to the right of the vertical 0-axis represent an improvement in word error rate with model Mix-Reverb+RealNoise. Similarly, all points above the horizontal 0-axis are an improvement with 3-fold. The dashed diagonal axis marks the transition where both models are equally good. If a point is above the line, the relative improvement is greater for 3-fold. Conversely, the improvement for Mix-Reverb+RealNoise is greater if the point is below the line. The square marks both models' overall test set word error rate improvements.

We are also interested in the effect of increasing the training data of 3-fold compared to Mix-Reverb+RealNoise. For these two models, we further analyze the relative improvements in Figure 4.9. The diagram shows no severe differences between Mix-Reverb+RealNoise and 3-fold since almost all results are close to the diagonal axis. Therefore, both models seem to improve similar acoustic aspects in recognition. For many interviews, 3-fold shows better recognition performance than Mix-Reverb+RealNoise. This is especially true for relatively high improvements in the upper right quadrant above 10 % relative. In this area, 3-fold is almost always at an advantage. For the other interviews, no model significantly dominates. Mix-Reverb+RealNoise is better for some interviews, 3-fold for others. This indicates that 3-fold plays out its advantage of more training data for acoustically more challenging interviews.

Preliminary Conclusion and Consideration of Required Training Time

Overall, the 3-fold model is the most robust model in this set of experiments and, thus, to be chosen for future applications. It shows the best generalization for out-of-domain data and, in addition, substantially improves the recognition for in-domain broadcast speech recordings. Therefore, we use this model for subsequent experiments and applications. However, it must be taken into account that training the 3-fold model is computationally expensive and slow due to the large amount of training data.

Training of the neural network is performed on overall 9000 hours of speech (3-fold speed perturbation and 3-fold noise and reverberation data augmentation) with features extracted every 10 milliseconds. This results in approximately $3.24 \cdot 10^9$ 300-dimensional inputs per epoch. However, neural network training is not the only issue. The 9000 hours of speech are time-aligned with bootstrap-trained GMM-HMM systems in advance.

Overall, on our department's GPU cluster at the Fraunhofer IAIS with a *GeForce GTX TITAN X* GPU node and a CPU node (two 12-core CPUs with 2.4 GHz), training of the 3-fold model took 66 days, i.e., more than two months. Of this, 23 days were spent on the GMM-HMM bootstrap training, feature extraction, and data alignment. Overall, 43 days were required for the GPU LF-MMI training of the TDNN-LSTM neural network. Therefore, comparative experiments with similarly large training data can only be performed in the present work to a minimal extent.

Table 4.12: Multi-condition bootstrap training (MB) compared to clean bootstrap training (CB). The reported results are word error rates (in percent) achieved by the final LF-MMI–TDNN-LSTM (with per-frame dropout). The best result between CB and MB is highlighted pair-wise for each Mix model.

		Mix-Reverb +WhiteNoise		**Mix-Reverb +BothNoises**		**Mix-Reverb +RealNoise**	
Test Set	**Clean**	**MB**	**CB**	**MB**	**CB**	**MB**	**CB**
GerTV Dev Set	13.8	**13.9**	**13.9**	**13.6**	14.2	**13.9**	**13.9**
DiSCo Average	12.8	**12.4**	12.9	**12.5**	12.7	**12.3**	12.8
Planned Clean	9.0	**8.8**	9.2	9.2	**9.0**	**8.9**	9.2
Planned Mix	12.0	**11.2**	11.6	**11.2**	11.7	**11.3**	11.6
Spontaneous Clean	10.6	**10.7**	10.8	10.5	**10.4**	**10.4**	10.8
Spontaneous Mix	19.7	**19.0**	20.2	**19.0**	19.8	**18.7**	19.5
Ger. Broadcast 2016	12.2	12.1	**11.6**	**12.0**	12.1	**12.3**	12.6
Challenging Broadcast	21.2	**20.4**	20.5	**20.5**	**20.5**	**20.2**	21.2
Oral History	34.2	**30.3**	31.0	**29.6**	30.7	**29.5**	31.0
Interaction	65.7	**50.8**	52.4	**49.9**	51.3	**49.8**	51.7
Spoken QALD-7	20.9	19.2	**19.1**	**19.0**	19.1	**18.6**	20.0

Ablation Study: Influence of Multi-Condition Training in Bootstrapping

In Table 4.12, we report the results of the ablation study on the influence of data augmentation in bootstrap training of the GMM-HMM models on the final LF-MMI acoustic model with fixed training data size. The experiments were conducted for all three configurations using reverberation and noises, as multi-condition training has the greatest influence with these setups. We use the phonetic decision tree and GMM clustering from the clean-trained baseline. The alignment for LF-MMI training is performed on the data augmented data sets using the clean-trained Triphone 4 model (cf. Table 4.7).

Most test sets have a better word error rate with the proposed approach than with clean-trained bootstrapping. In setup Mix-Reverb+WhiteNoise, only German Broadcast 2016 and Spoken QALD-7 have lower word error rates with the clean-trained bootstrap. Interestingly, the improvement is quite substantial for German Broadcast 2016, without evidently recognizable reasons. In Mix-Reverb-+BothNoises, which has a somewhat higher proportion of noise than Mix-Reverb-+WhiteNoise, the two clean DiSCo subsets have slightly better word error rates

Table 4.13: Results of multi-condition-training achieved by the CE-LSTM model using the 128-hour subset of the GerTV1000h corpus. Results are word error rates in percent.

Test Set	Clean	Mix-Reverb	Mix-Reverb +WhiteNoise	Mix-Reverb +BothNoises	Mix-Reverb +RealNoise
GerTV Dev Set	**16.6**	17.1	17.3	17.4	17.3
DiSCo Average	**17.2**	17.6	17.7	18.0	17.6
Planned Clean	**12.1**	12.4	12.7	12.8	12.4
Planned Mix	**16.7**	16.9	16.8	17.3	16.7
Spontaneous Clean	**13.9**	14.4	14.5	15.1	14.6
Spontaneous Mix	**26.0**	27.0	26.7	26.8	26.5
German Broadcast 2016	15.1	**14.9**	15.2	15.4	15.5
Oral History	44.1	40.0	39.5	39.1	**38.9**
Interaction	78.6	65.6	**62.5**	62.6	63.4

with the clean-trained bootstrap. For Mix-Reverb+RealNoise, the setup with multi-condition bootstrapping is better on all test sets.

Generally, applying multi-condition training data in bootstrapping leads to more robust LF-MMI acoustic models. In particular, the improvement is quite substantial for acoustically challenging tasks, such as our oral history use case. Therefore, we consider this approach useful for our use case, although re-performing bootstrapping for each multi-condition setup increases training time.

Ablation Study: Proposed Data Augmentation with Different Model, Training Criterion, and Training Data

In Table 4.13, we present the results of the ablation study with a different acoustic model architecture (LSTM instead of TDNN-LSTM), a different training criterion (CE instead of LF-MMI), and different training data (128 instead of 1000 hours). The models for this experiment were trained before conducting the LF-MMI experiments with 1000 hours of training data and serve here as an ablation study. This experiment intends to expose to which extends the observed improvements of multi-condition training for in-domain and out-of-domain data are a property

of the approach itself or whether it is dependent on the model or training data. The results were first published in [Gref et al., 2018a].

In contrast to the LF-MMI–TDNN-LSTM model experiments on the entire training data at the beginning of this section, multi-condition training does not improve the broadcast domain. Subsequent, very time-consuming ablation studies would be necessary to investigate the exact contribution of each of the three changed parameters to this result. However, since our primary objective is the investigation of oral history interviews and domain mismatch to the training data, we are content with this investigation and leave this to future work.

We observe a substantial improvement for the two out-of-domain test sets, similar to the LF-MMI–TDNN-LSTM experiments with 1000 hours of training data. In particular, for oral history, the word error rate substantially improves with all multi-condition setups. Moreover, we observe similar trends of the different multi-condition setups on the oral history test set: modeling room reverberation in Mix-Reverb achieves more than 9 % relative improvement to the baseline. Also, modeling additive noises further improves the recognition performance. Using real-world noises and the proposed random noise overlapping in Mix-Reverb+RealNoise achieves the best results. These results indicate that, in principle, the proposed data augmentation for multi-condition training improves ASR performance for oral history interviews with different model types and training data sets.

4.4.5 Improved 3-fold Acoustic Model and Language Model Comparison

Over the course of the present research work, we improved the 3-fold model from the previous experiment (*3-fold v1*) in two iterations. These models serve as improved baselines for different experiments in the following chapters. The results for all three variants of the 3-fold models are summarized in Table 4.14. We report the word error rates using the default and large language models (cf. Section 3.5), respectively.

First, we point out the substantially and consistently better results of the large language model on Spoken QALD-7, Challenging Broadcast, and German Broadcast 2016. The substantial improvements of the large language model result from the better perplexity of this language model for these data sets, cf. Table 3.8. As described in Section 3.5, the large language model's training data size and vocabulary are multitudes larger than the default's. It models typical formulations in the broadcast domain better and includes considerably more proper names and other entities. In particular, it contains substantially more entities frequently occurring in the Spoken QALD-7 test set, which causes substantial improvement for this domain. For Oral History, the perplexity of both language models is very simi-

Table 4.14: Comparison of the original and improved 3-fold LF-MMI–TDNN-LSTM acoustic models. Results are reported as word error rates in percent for the default and large language model. The best result per test set is highlighted pair-wise for each language model.

Language Model:	**Default LM**			**Large LM**		
3-fold Acoustic Model:	**v1**	**v1.1**	**v2**	**v1**	**v1.1**	**v2**
GerTV Dev Set	13.7	**13.6**	**13.6**	12.9	12.9	**12.8**
DiSCo Average	**11.8**	11.9	12.0	**11.9**	12.2	12.1
Planned Clean	9.0	9.0	**8.9**	**8.8**	9.0	**8.8**
Planned Mix	10.8	11.0	**10.6**	9.9	9.9	**9.8**
Spontaneous Clean	**9.9**	10.0	10.7	**10.7**	11.1	11.1
Spontaneous Mix	**17.5**	17.6	17.9	**18.4**	18.6	18.8
German Broadcast 2016	11.5	11.7	**10.9**	9.7	9.9	**9.2**
Challenging Broadcast	20.1	19.7	**19.4**	17.8	17.4	**17.2**
Oral History	28.2	27.7	**26.6**	27.2	27.1	**26.0**
Interaction	47.8	48.2	**47.6**	50.9	51.2	**49.6**
Spoken QALD-7	**18.3**	19.0	**18.3**	14.4	14.8	**13.6**

lar. Nevertheless, the large language model consistently performs better on Oral history than the default—probably due to the better OOV rate, cf. Table 3.9.

The first improvement of the 3-fold acoustic model is *3-fold v1.1*, which finished training in mid-2019. For the training of this model, more heterogeneous data for the i-vector extractor training and a slightly adjusted learning rate scheduling for the acoustic model neural network was used. Furthermore, we adjusted the subset sizes for the bootstrap model training. Instead of fixed numbers of segments (cf. Section 4.4.3), fixed relative ratios of segments from the entire data set are used. This enables the training routine to handle differently sized training data automatically. The 3-fold v1.1 acoustic model performs a little worse in the common broadcast domain but better in our target domain Oral History and on Challenging Broadcast. This is true regardless of the language model chosen in our experiments. The 3-fold v1.1 was used as a baseline in [Gref et al., 2020] for cross-lingual domain adaptation experiments presented in Chapter 6.

Towards the end of the presented research work, a further improved version of the 3-fold model, *3fold v2*, was trained and released. This model is more robust than 3-fold v1 and 3fold v1.1 for almost all of our test sets. The improved 3-fold v2 model achieves better or equally good recognition performance on all test sets except for the DiSCo Spontaneous subsets. Overall, the results are consis-

tent with the different acoustic models using both language models. In particular, for Oral History, this model substantially improves the recognition performance and achieves a 26.6 % word error rate with the default language model. A detailed analysis of the recognition results on all test sets reveals that with 3-fold v2, we observe an increased number of deletions but significantly reduced substitutions and insertions. These substantial improvements are achieved by increasing the amount and diversity of the training data, utilizing a large commercial noise database, and using more training data in bootstrapping.

In detail, for 3-fold v2, we obtained 350 hours of additional German speech data for training from different domains, such as podcasts, various interviews recorded in different acoustic conditions, read speech, and political speeches. Overall, 1345 hours of source data (*GerTV1345h*) were used for training. As for the original 3-fold model, the training data was increased 9-fold with 3-fold speed perturbation and 3-fold noise and reverberation data augmentation.

Additionally, we purchased a large collection with different sounds and ambient noises recorded in various environments that we use as the real-world noises in the noise and reverberation data augmentation. Overall, for 3-fold v2, we used 320 hours of noises instead of the small 14.5-hour noise set used previously for 3-fold v1 and 3-fold v1.1. We omit the random noise overlap approach from the data augmentation for this model since the amount of speech and noise are in a reasonable ratio.

The 3-fold v2 model has been the standard acoustic model of the Fraunhofer IAIS Audio Mining system since its release in 2020. It is deployed in various client systems, including the ARD, where it transcribes the media library for journalists and archives. The model was first considered in the literature in [Gref et al., 2022b] and used for adaptation experiments for oral history presented in Section 5.5.

For academic purposes, the 3-fold v2 model can be used for free with a limited monthly contingent as part of the *BAS Speech Science Web Services* [Kisler et al., 2016].[1] The version used for the web service uses an updated and improved version of the large language model.

4.4.6 Summary and Conclusion

The presented experiments studied data augmentation for multi-condition training for DNN-HMM acoustic models to reduce the domain mismatch between the broadcast training and oral history interviews. We discussed the acoustic conditions of the studied oral history interviews and identified additive noises and, in particular, reverberation as the primary challenges of oral history interviews for the acoustic model.

[1] https://clarin.phonetik.uni-muenchen.de/BASWebServices

We proposed and studied the application of noise and reverberation data augmentation for multi-condition training. Additionally, we proposed multi-condition training during the GMM-HMM bootstrap training and studied the influence on the final LF-MMI acoustic model.

The experiments show that noise and reverberation data augmentation improves the robustness of the studied LF-MMI–TDNN-LSTM model not only for acoustic conditions of oral history interviews but also for the broadcast domain. This effect becomes stronger when the training data size is increased—but can also be observed without increasing the data size.

We performed experiments with different setups to investigate different properties of multi-condition training. We showed that real-world noise leads to better generalization than artificial AWGN, even if a comparable small noise database is used. We showed that the improvements from multi-condition training for the oral history use case do not depend on acoustic model architecture, the training criterion, and the training data size.

By using numerous test sets with different properties, we were able to show how the training affects different domains. Overall, we observe a remarkable generalization and a decreased domain mismatch of the acoustic model with the proposed multi-condition training. For the studied oral history use case, we achieved a relative improvement of 17.6 % with 3-fold multi-condition training compared to the clean-trained baseline.

4.5 Study: Speech Enhancement for Robust Speech Recognition

In the overview on robust speech recognition given in Section 4.2, we discussed the two main recent directions: multi-condition training and speech enhancement. Due to promising results in the literature and some of our initial experiments, we focused mainly on multi-condition training in this chapter.

Nevertheless, speech enhancement is an important component in many robust ASR applications. Some applications, e.g., Du et al. [2018] and Kanda et al. [2018], combined speech enhancement with multi-condition training. In this section, we study promising speech enhancement approaches for the oral history use case.

4.5.1 Experimental Setup

Speech enhancement algorithms are often developed and used for specific types of distortions. Thus, we again focus on acoustic distortions, particularly noise and reverberation. In our experiments, we study:

Table 4.15: Comparison of different speech enhancement approaches for the Oral History test set with the clean-trained and robust 3-fold (v1) TDNN-LSTM–LF-MMI model. Results are reported as word error rates in percentage.

		Acoustic Model	
Approach	**Implementation**	**Clean**	**3-fold**
Original	no enhancement	34.2	28.2
Adaptive Filtering	Hirsch [2014]	35.1	29.0
DNN Enhancement	Hirsch and Gref [2017]	57.4	46.0
Multi-Band Compression	Jaeger et al. [2019]	69.6	52.6
WPE	Drude et al. [2018]	**33.5**	**27.9**

- a classical adaptive filtering implementation by Hirsch [2014] with cepstral smoothing [Breithaupt et al., 2007].
- a DNN-based speech enhancement from a preliminary work [Hirsch and Gref, 2017] trained on Aurora4 corpus [Parihar and Picone, 2002] for noise reduction and different microphone qualities.
- a speech enhancement approach based on multi-band compression with self-regulation based on objective speech intelligibility estimation by Jaeger et al. [2019].
- blind speech dereverberation with linear prediction called *weighted prediction error* (*WPE*) proposed by Nakatani et al. [2008] and extend in recent years. For the experiment, we use the NARA-WPE implementation by Drude et al. [2018].

The first three approaches mainly focus on additive noises and noise reduction. The second, DNN-based approach is additionally trained to compensate for different microphone qualities, as in the Aurora4 training data. The last approach focuses on dereverberation.

For decoding, we use the default broadcast language model. We use the clean-trained and robust 3-fold (v1) model from the previous section for evaluation. The evaluation is performed on the Oral History test set.

4.5.2 Results and Discussion

The results are presented in Table 4.15 for the clean-trained and the 3fold (v1) LF-MMI–TDNN-LSTM models. Although all approaches improve perceived audio

quality, all except one approach resulted in no improvement. These approaches even substantially decreased the recognition performance when applied as preprocessing to the ASR system.

We observe a slight deterioration of the average recognition performance for classical adaptive filtering. However, for DNN-based enhancement and multi-band compression, the degradation in recognition performance is substantial. Both approaches are designed to improve auditory quality and intelligibility for humans. The results show that these auditory improvements do not correlate with recognition performance for our ASR system. Furthermore, DNN filtering trained on Aurora4 data shows poor generalization or domain overfitting for the oral history interviews. Artifacts in the spectrogram negatively affect recognition performance.

However, blind dereverberation using the weighted prediction error results in a consistent improvement, both for the clean and the robust acoustic model. The improvement by WPE is a bit more substantial for the clean-train model but not as good as the multi-condition training without speech enhancement, cf. Table 4.10. For the robust 3-fold model, WPE adds a slight improvement. Thus, this approach should be reconsidered as an addition to the final acoustic model in the final system.

4.6 Summary and Contributions

4.6.1 Summary

In this chapter, we conducted various experiments and detailed investigations on the robustness of acoustic models for oral history interviews. After giving an overview of the general approaches of robust speech recognition, we proceeded in three subsequent studies with several experiments. For all investigations, we always used a multitude of test sets from different domains to obtain reliable estimates of the real-world performance and avoid the selection of models that suffer from domain overfitting.

In the first study in this chapter, we compared different acoustic models and training criteria that were state-of-the-art in 2017. The goal was to identify the acoustic model that shows the best initial results, robustness, and generalization properties for different domains for further investigations.

We selected four different hybrid acoustic model candidates from the literature and compared them using two different training data sizes. We used a moderate training data size with 128 hours of broadcast speech and a large training data size with 1000 hours. A TDNN-LSTM architecture trained with the LF-MMI criterion and per-frame dropout achieved the best performance and highest generalization for the different domains with both training set sizes. The best system in this

study achieves a 34.3 % word error rate on our oral history test set, outperforming the previous 55 % word error rate baseline by a substantial margin.

In addition, we examined the model's susceptibility to different speech segmentation granularities. We found that LF-MMI trained models are more susceptible to long segments than cross-entropy trained models. Long segments lead to a substantial deterioration in recognition performance for the studied LF-MMI models. This is considered in the processing workflow of the Audio Mining system that performs utterance segmentation.

In the second study in this chapter, we investigated multi-condition training with data augmentation to improve the LF-MMI–TDNN-LSTM model's robustness for the acoustic conditions in oral history interviews. First, we identified room reverberation as one of the primary challenges of our oral history interviews through qualitative analysis of spectrograms from different interviews. We proposed noise and reverberation data augmentation with specific details to improve robustness. In particular, we combine clean and augmented speech to perform multi-condition instead of match-condition training, which would not be feasible for our oral history interviews. We further propose using multi-condition training in the entire GMM-HMM bootstrap instead of only in the neural network acoustic model stage to improve robustness further. We proposed an approach for noise generation for real-world noise to compensate for a small real-world noise database. Several experiments investigated the influence of bootstrap training, acoustic model architecture, training criterion, and training data size on the robustness of multi-condition training with data augmentation.

We explored different combinations of acoustic conditions and identified reverberation with real-world noise as the most robust combination in our experiments. Even without increasing the training data set, our multi-condition training improves robustness not only for the oral history domain but also for the broadcast, interaction, and speech assistant domains. With a 3-fold increase in training data size using the proposed data augmentation, an even more substantial improvement in robustness is achieved for all domains. This model achieves a 28.2 % word error rate on the oral history test set. As this model performs substantially better in the broadcast domain than a comparable clean-trained model, as a by-product, this model has become the standard acoustic model for the Fraunhofer IAIS Audio Mining system and is deployed and in daily use at clients such as ARD.

The last study investigates different speech enhancement methods for acoustic robustness on oral history interviews. We explored different approaches for the clean-trained and robust model from the previous study. The common approach for blind dereverberation using the weighted prediction error is one of the few effective approaches for oral history interviews. It does not achieve the same robustness as

the multi-condition trained acoustic models. However, it can additionally improve the word error rate of the robust model to 27.9 % on the oral history test set.

In summary, multi-condition training with noise and reverberation data augmentation results in substantial improvements in terms of the word error rate for oral history interviews, other out-of-domain data, and also improves recognition for in-domain broadcast training data. In conclusion, the results indicate that this training approach reduces the domain mismatch and improves generalization as it not only improves the acoustic robustness of the model but, also also improves recognition for spontaneous speech to a certain degree. However, oral history interviews still remain challenging as we still observe a wide range of recognition accuracies, i.e., 14.8–52.8 % word error rates in the interviews. This indicates that many of the previously discussed challenges in Section 3.3.2 still prevail. Therefore, in the following chapters, we aim to further improve the acoustic model to conditions not modeled by the data augmentation and the interviewees' speech characteristics by adapting the acoustic model.

4.6.2 List of Contributions

List of scientific contributions in this chapter:

- Noise and reverberation data augmentation was proposed and studied to improve the real-world performance of LF-MMI acoustic models for oral history interviews and other domains with unseen conditions by reducing the domain mismatch and improving acoustic robustness.
- A selection of current hybrid acoustic models was explored and compared in general performance (for in-domain data) and domain mismatch (for out-of-domain data).
- The influence of fine and coarse segmentation of speech was studied for decoding with cross-entropy and LF-MMI acoustic models.
- The reverberation caused by small and medium-sized rooms was identified as one of the primary challenges for acoustic models of German oral history interview recordings.
- Different speech enhancements methods were investigated to improve acoustic robustness for oral history interviews for clean-trained and robust LF-MMI acoustic models.

5 Acoustic Model Adaptation Using Transfer Learning

In the previous chapter, we improved the robustness of acoustic models against challenging acoustic recording conditions that are common in many oral history interviews. We proposed and investigated noise and reverberation data augmentation for multi-condition training of the acoustic model that substantially improved the recognition performance not only for oral history interviews but also for speech recordings from the broadcast and other domains.

A limitation of these experiments is the tremendous amount of time required to train hybrid acoustic models from scratch on several thousand hours of annotated speech. Further improvement can be achieved with the further enlargement of the training data by the proposed and additional data augmentation. However, training a single acoustic model with this amount of data in our infrastructure would take several months. This training time is not feasible for research and real-world applications in new domains.

Moreover, while data augmentation is quite successful in overcoming a mismatch in acoustic conditions between desired applications and training data, it is limited to acoustic distortions that can be artificially created. The remaining challenges in oral history interviews, such as peculiarities in the way of speaking, spontaneous speech, and dialects, cannot be modeled with the approach. For example, as shown in the joint investigation [Gorisch, Gref, and Schmidt, 2020] with the *Leibniz Institute for the German Language* (*IDS*)[1], the dialect has a substantial influence on the recognition performance of our system. It may increase the word error rate up to 60 %.

In this chapter, we investigate the domain adaptation of the acoustic model using fine-tuning to overcome these limitations and further improve the recognition performance. In particular, we examine possible domain mismatches due to the adaptation to obtain a realistic estimate of the real-world performance and robustness of the adapted models. In total, we present three independent, subsequent studies in this chapter. These studies were conducted some years apart from each other, when additional oral history data for adaptation and evaluation became available, enabling us to study different aspects and facets of the adaptation in more detail.

[1] https://www.ids-mannheim.de

In the first of the three studies, we investigate the general idea of fine-tuning for the domain adaptation of the acoustic model with the 3.5-hour oral history data set using a leave-one-speaker-out cross-validation approach in Section 5.3. In particular, we investigate the extent to which the data augmentation from the previous chapter complements fine-tuning and the respective influence of the two methods on the recognition.

We investigate automatic transcript alignment in the second study in Section 5.4 as a method to generate data for adaptation semi-automatically to overcome the limitation of a lack of suitable annotated data for adaptation. We also investigate the impact of different training data sets and learning rates on potential domain overfitting.

In the third study in Section 5.5, we investigate domain overfitting within the oral history domain with additional oral history adaptation and test data from a different archive. These new interviews were recorded in more recent years and thus have substantially better recording conditions. This new data allows us to study the influence of the acoustic conditions and the other challenges of oral history interviews for speech recognition adaptation in more detail.

In Section 5.6, we summarize the three studies' findings and contributions.

5.1 Thesis Author Contribution

Parts of this chapter are covered in the publications:

> Michael Gref, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Two-staged acoustic modeling adaption for robust speech recognition by the example of German oral history interviews. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 796–801, 2019. doi:10.1109/ICME.2019.00142

> Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*, 2022b. URL https://arxiv.org/abs/2201.06841

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author.

The experiments of the first adaptation study summarized and extended in Section 5.3 are based on [Gref et al., 2019]. The experiments of the third adaptation study summarized and extended in Section 5.5 are based on [Gref et al., 2022b]. Additional experiments and investigations, e.g., on the influence of the learning rate and adaptation data size for both studies were conducted in the presented research work.

5.2 Related Work

In recent years, transfer learning for acoustic model adaptation has raised attention, in particular, for under-resourced language and domain tasks. Transfer learning is an approach to improve generalization and performance by transferring knowledge of a model trained in one domain to train a model in another related domain [Goodfellow et al., 2016]. It is advantageous in scenarios where only little training data is available for the main task, but a large amount of annotated data is available for a similar or related task.

Wang and Zheng [2015] give a detailed overview of transfer learning in speech and language processing. Ghahremani et al. [2017] investigated transfer learning using weight transfer for LF-MMI models for several well-known English speech recognition tasks. Transfer learning was also studied for end-to-end speech recognition systems utilizing end-to-end systems' particularities, such as Ueno et al. [2018] fine-tuning the encoder of attention-based models using Japanese corpora.

However, most works in automatic speech recognition, such as the aforementioned, studied transfer learning with a much greater amount of annotated speech than is available in the oral history task. In addition, most works on domain adaptation in ASR focus on either data augmentation or transfer learning, usually to address a particular task or challenge, such as robustness to noise and not the holistic robustness of an acoustic model.

5.3 Study: Two-Staged Acoustic Modeling Domain Adaptation

In this first of three studies on transfer learning for the acoustic model, we investigate the combination of acoustic model fine-tuning with data augmentation—which can be understood as the adaptation of training data—in a two-staged approach. We are particularly investigating how both approaches relate to each other in terms of domain adaptation and robustness improvements and whether the two methods complement each other for real-world application. This study is based on [Gref et al., 2019].

5.3.1 Proposed Two-Staged Acoustic Modeling Adaptation

We aim to improve the performance of robust acoustic models by performing a two-staged acoustic modeling adaptation using very little training data from the target domain—German oral history interviews in our study. An overview of the proposed method is given in Figure 5.1.

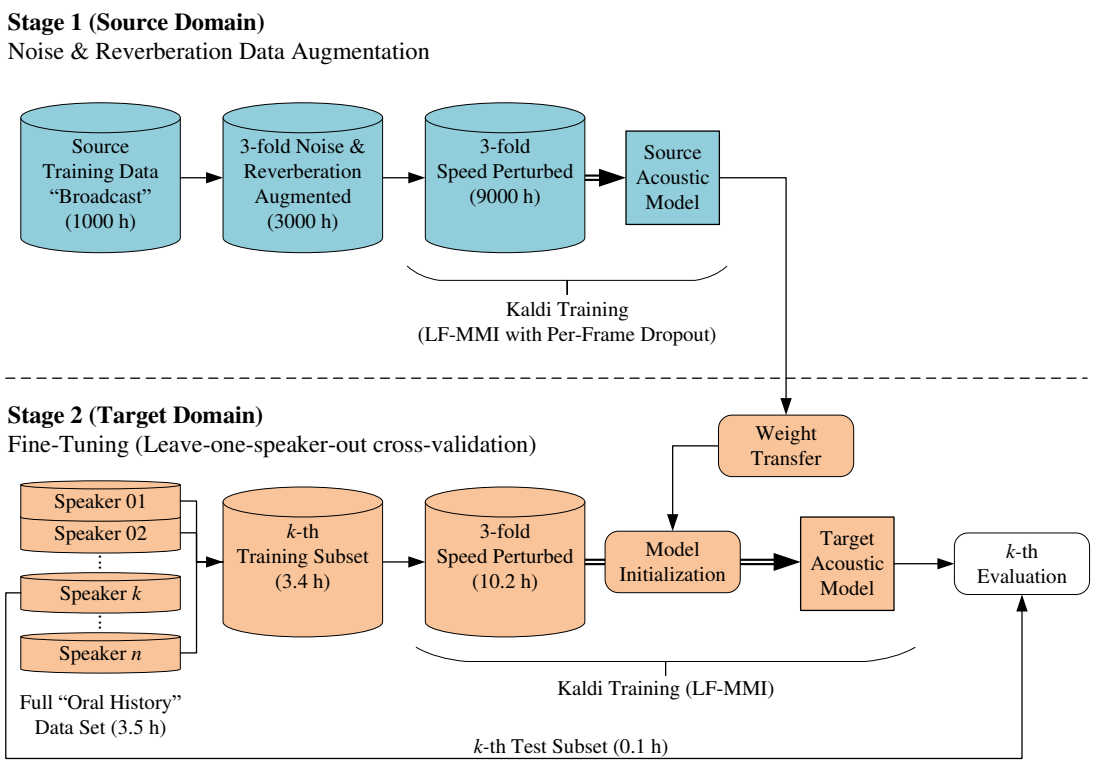


Figure 5.1: Proposed approach for two-staged acoustic model adaptation and evaluation with leave-one-speaker-out cross-validation. Noise and reverberation data augmentation is applied in Stage 1 to adapt the training data to the acoustic conditions of oral history interviews and increase overall robustness. In Stage 2, transfer learning is applied to tackle remaining challenges such as interviewees' characteristics and spontaneous speech.

The first stage of the approach is based on the 3-fold model of Section 4.4. We use multi-condition training with noise and reverberation data augmentation to decrease the acoustic domain mismatch between conventional clean broadcast training data and oral history interviews. This has been proven to significantly increase the performance of speech recognition systems on German oral history interviews and improve overall robustness.

The second stage applies transfer learning to tackle the remaining acoustic challenges and interviewees' speech characteristics in the target domain that could not be synthesized in the first stage—such as spontaneous speech, dialectics, and pronunciations. The transfer learning in Stage 2 is inspired by Ghahremani et al. [2017]. In our setup, however, a full weight transfer of the entire source model for initialization of the target model is applied without layer freezing. In particular, the output layer is not replaced in contrast to some other transfer learning approaches in speech recognition since we use the same set of phonemes and the

Table 5.1: Training parameters in both stages of the two-staged acoustic model adaptation.

Parameter	Full Training (Stage 1)	Fine-Tuning (Stage 2)
Initial learning rate	1e-3	1e-6
Final learning rate	1e-4	1e-7
Per-Frame Dropout	yes	no

same decision tree, both in the source and target scenario. In the transfer learning stage, the i-vector extractor of the model trained in Stage 1 is utilized without adaptation.

In both stages, we apply speed-perturbation proposed by Ko et al. [2015] on the entire training data to further increase the amount of data three-fold before neural network training. We consider this a part of the standard Kaldi training routines for LF-MMI systems.

The neural network training routine in Stage 2 is almost equal to Stage 1. An overview of the parameters that are different in the transfer learning stage is given in Table 5.1. The values for Stage 1 are our default values for acoustic model training. The values for Stage 2 are obtained in preliminary experiments. In Stage 1, we apply per-frame dropout according to Cheng et al. [2017]. As observed in the experiments in Section 4.3, per-frame dropout improved the generalization of the LF-MMI trained models with 128 and 1000 hours of training data. However, our preliminary experiments with transfer learning showed that dropout reduces performance when training on small data sets. Therefore, the training in Stage 2 is performed without per-frame dropout. The training is performed for four epochs in both stages with a decaying learning rate with fixed decay. The initial and final learning rate in the second stage is lower than in the first stage due to the significantly reduced training data size.

5.3.2 Leave-One-Speaker-Out Cross-Validation

We lack suitable training data for the oral history domain at the time of conducting the study and only have a 3.5-hour data set with 35 different speakers that we usually use for testing only. To study transfer learning, we apply a strategy similar to k-fold cross-validation: *leave-one-speaker-out cross-validation*.

As for k-fold cross-validation, we partitioned the data set in k subsets and iterated training and validation with varying subsets. However, this is not done randomly but according to speakers. Each of the k subsets consists of precisely

one speaker. We loop k times over the data subsets and keep one speaker out of the training set for validation and train a model the remaining $k-1$ speakers, as illustrated in Figure 5.1. This way, we run k experiments in Stage 2 and evaluate each trained model on the speaker absent in the training data.

5.3.3 Experimental Setup

The experiments are carried out using the Kaldi ASR toolkit. We use the same TDNN-LSTM topology with LF-MMI training we studied in Chapter 4.

Training Data

For training in Stage 1, we use the 3-fold (v1) acoustic model as presented in Chapter 4. In summary, we created the following two artificially distorted versions of the 1000 hour GerTV1000h source training data. We combined them with the original (*clean*) set to create a 3000-hour multi-condition source training set:

- **Reverb**: All signals are convolved according to Equation 4.2 with randomly selected room impulse responses of small or medium-sized rooms from our data collection. No noise is applied here.
- **Reverb+RealNoise**: Similar to **Reverb** but randomly selected noise recordings from our database are included according to Equation 4.1, applying a random signal-to-noise ratio between 10 and 20 dB.

This broadcast training data can be considered out of domain for the oral history scenario. As presented in Section 3.3.2 and further analyzed in Section 4.4.1, the broadcast recordings severely differ from oral history interviews in many different aspects, such as recording technology, audio signal quality, and speech characteristics.

We use our 3.5-hour oral history test set with 35 different interviewees for the leave-one-speaker-out cross-validation experiments. This data size is substantially smaller than the adaptation data size in other transfer learning works in the field of ASR. As in all prior experiments, we use the additional broadcast and other-domain test sets for evaluation to assure robustness, detect domain overfitting, and obtain a reliable estimate of the real-world performance for unseen data.

Decoding

In contrast to all prior and subsequent experiments, in this study, we explicitly consider the speaker identity to be known for each segment in the oral history set. This is required for the k-fold cross-validation. Additionally, considering speaker

Table 5.2: Comparison of the best language model weights (LMWT) on models from the multi-condition experiments. The primary LF-MMI–TDNN-LSTM models from Table 4.10 are considered.

Model	GerTV Dev	Oral History (Test)
Clean	8	8
Mix-Reverb	8	8
Mix-Reverb+WhiteNoise	9	8
Mix-Reverb+BothNoises	8	8
Mix-Reverb+RealNoise	8	8
3-fold (v1)	8	8

identities to be known can also affect the decoding. In particular, the i-vector extraction can be performed across multiple segments instead of segment-wise to obtain more robust speaker embeddings and improve the acoustic model's speaker adaptation, cf. Xue et al. [2014]. We refer to this as *speaker-aware* decoding in contrast to our default *speaker-unaware* decoding.

In [Gref et al., 2019], we only evaluated and reported the systems with speaker-aware decoding of the oral history test set. As an extension in the presented research, we additionally perform and report speaker-unaware decoding. This makes the results comparable to the previous and following experiments and better reflect the expected real-world performance in the Audio Mining system, where we decode each segment independently, i.e., in a speaker-unaware manner. Additionally, by comparing speaker-aware and speaker-unaware decoding, we gain insights into the influence on our models.

We use the same default 5-gram broadcast language model and G2P-based pronunciation lexicon as in the previous experiments. However, in this study, we do not adjust the language model weight for decoding for each model. As discussed in Section 4.4.4, adjusting each model solely to the GerTV development set may not necessarily lead to the best results. Therefore, determining a specific language model weight for each of the 35 leave-one-speaker-out experiments based on the GerTV development set involves the risk of distorting the recognition results due to high statistical noise. Instead, we use a fixed language-model weight of eight and do not adjust this parameter for the adaptation experiments. This fixed value also seems to be the best configuration for most models, including the models from the previous chapter as summarized in Table 5.2.

Ablation Studies

In this study, we are particularly interested in studying the differences, similarities, and combination of transfer learning and data augmentation in terms of domain adaptation, robustness improvements, and real-world performance. Therefore, we perform ablation studies to investigate the individual influence of the two methods on the final performance by comparing four different models:

1. **Proposed Approach**: Applying both Stage 1 and Stage 2, i.e., apply transfer learning with 3.5 hours oral history data on the 3-fold trained robust acoustic model.
2. **Stage 1 Only (Data Augmentation)**: Evaluating the performance of the source model trained in Stage 1 using the 3-fold noise and reverberation data augmentation.
3. **Stage 2 Only (Transfer Learning)**: Applying transfer learning on a clean-trained baseline model (without noise and reverberation data augmentation).
4. **Baseline (Clean)**: Clean-trained baseline model with 1000 hours of broadcast speech without transfer learning and noise and reverberation data augmentation.

We had to re-train the clean baseline model with an updated version of the Kaldi ASR toolkit to apply transfer learning approaches for comparison experiments. Due to the default non-deterministic dithering in MFCC feature extraction in Kaldi, the baseline model results slightly vary compared to Section 4.4. However, the deviations are within the range of decimal places and do not influence the conclusions drawn.

Evaluation Metrics for Leave-One-Speaker-Out Cross-Validation

As usual, we use the overall word error rate of entire test sets defined according to Equation 3.1 to report the results of the experiments. However, for leave-one-speaker-out cross-validation of the fine-tuned models, the evaluation is performed on speaker subsets from the oral history data set. We report two values to assess the cross-validation better and make the results comparable to previous and following word error rates reported on the oral history data set.

First, we report an accumulated word error rate of the oral history test set. We sum up the number of word errors and number of words in the reference for each of the 35 different interviews in the 35 separate leave-one-out experiments. This accumulated word error rate is the quotient of the sum of word errors and the sum of reference words. This corresponds to the conventional calculation of

word error rate on the entire test set. Thus, this word error rate is comparable to the results of previous and following experiments without leave-one-speaker-out cross-validation.

Additionally, we report the arithmetic mean and standard deviation of the individual word error rates of all 35 different experiments. The conventional (accumulated) word error rate and this *average of word error rates* are similar but vary slightly. In the default word error rate calculation, longer interviews have a stronger effect on the overall error rate since each interview has a slightly different length and number of words in the reference. Mathematically, the conventional word error rate can also be understood as the *weighted* arithmetic mean of word error rates of interviews, where each interview's word error rate is weighted with the number of words in the reference.

5.3.4 Results and Discussion

In the following, we present the results of the two-stage acoustic model adaptation experiments. We first report the results for the leave-one-speaker-out cross-validation using the oral history data set. We show that the proposed combination of fine-tuning and data augmentation improves the recognition for oral history interviews better than one of the approaches alone. Then we study a possible domain overfitting and real-world performance of the proposed method by evaluating the approach on multiple data sets from different domains.

Leave-One-Speaker-Out Cross-Validation

The results of the 35 leave-one-speaker-out cross-validation experiments using the 3.5-hour oral history data set are summarized in Table 5.3. We report results both for speaker-aware and speaker-unaware decoding.

The accumulated and average word error rates are very similar for all configurations, differing by a maximum of 0.2 percentage points in absolute terms. This indicates a very homogeneous distribution of the length of the different oral history interviews in our data set.

As expected, speaker-aware decoding yields better results than conventional, speaker-unaware decoding. The absolute difference is about 1.1–1.9 percentage points and roughly 4–6 % relative. The deviation is smaller for the more robust models *3-fold* and *3-fold Fine-Tuned* than for the less robust models. Thus, these models seem to cope slightly better with speaker-unaware decoding, which is advantageous for the application in our Audio Mining system.

Overall, the proposed two-stage approach shows the best performance for both decodings. By fine-tuning the 3-fold model, a relative improvement of 5.7 % (speaker-aware) and 5.8 % (speaker-unaware) is achieved. For the clean-trained

Table 5.3: Results of the leave-one-speaker-out cross-validation experiments for the two-staged acoustic model adaptation using the oral history data set. Results are reported for speaker-aware and (the default) speaker-unaware decoding, both as the accumulated (Acc.) test set word error rate and average (Avg.), i.e., arithmetic mean of the 35 interviews' word error rate ± standard deviation.

		Speaker-Aware		**Speaker-Unaware**	
Stages	**Model Name**	**Acc.**	**Avg.**	**Acc.**	**Avg.**
1+2	3-fold Fine-Tuned	25.5	25.4 ± 7.5	26.6	26.5 ± 7.5
1	3-fold	27.0	26.9 ± 7.6	28.2	28.1 ± 8.3
2	Clean Fine-Tuned	29.6	29.5 ± 9.2	31.5	31.4 ± 10.5
None	Clean	31.6	31.4 ± 10.0	33.4	33.2 ± 11.3

model, the relative improvement due to fine-tuning is 6.3 % (speaker-aware) and 5.6 % (speaker-unaware). These substantial improvements are remarkable since a comparatively small data set was used for the adaptation. However, the effect of fine-tuning the clean-trained baseline model on 3.4 hours is not sufficient to achieve the same performance of 3-fold data augmentation.

The fine-tuning of the acoustic model, but especially the data augmentation, also reduce the standard deviation of the word error rates. To further investigate the distribution of the word error rates of the 35 different interviews, the box plot diagrams of the word error rates with speaker-unaware decoding for each interview are shown in Figure 5.2.

The improvement due to fine-tuning has a particular impact on the interviews with higher word error rates. While the lower quartile only slightly changes due to the adaptation for Clean and 3-fold, the median and upper quartile for both models decrease with fine-tuning. Using the proposed two-stage approach, about 75 % of the interviews achieved a word error rate below or near 30 %. Except for one outlier, all interviews have a word error rate below or near 40 %. While half of the interviews have an error rate below 28.0 % for the 3-fold model, the median reduces to 25.6 % with fine-tuning. Overall, the box plots reveal that 3-fold data augmentation leads to more significant improvements for the oral history data. However, both the Clean and 3-fold models similarly benefit from adaptation through fine-tuning.

We report the results for speaker-unaware decoding as they reflect the expected real-world performance in Audio Mining and are comparable to the other experiments in the presented research work. The corresponding diagrams for speaker-aware decoding can be found in Appendix B.3. An additional evaluation of these

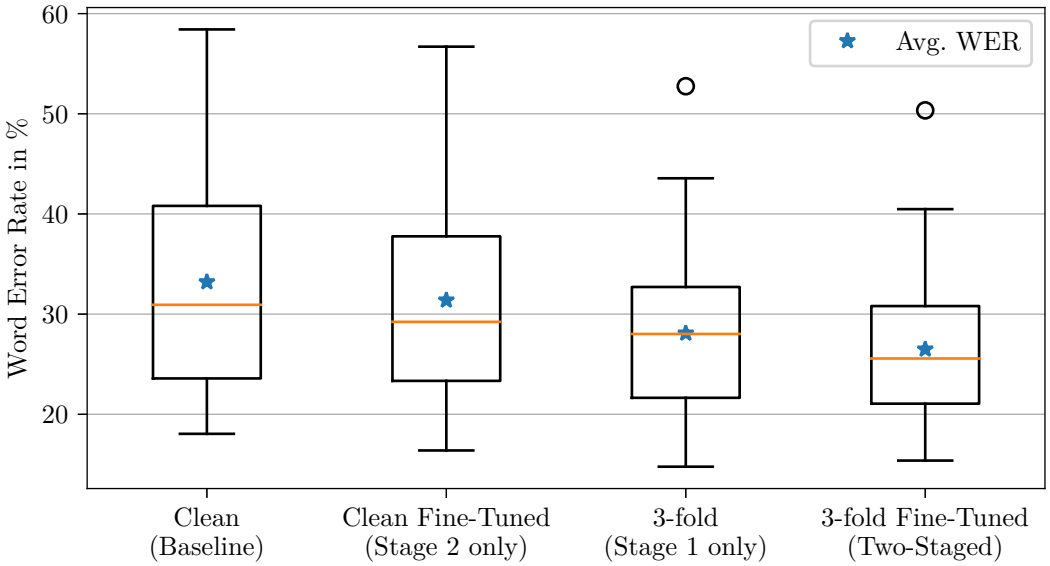


Figure 5.2: Box plot diagram of the word error rates of the 35 interviews for each model in the leave-one-speaker-out experiments with speaker-unaware decoding.

diagrams for speaker-aware decoding does not provide additional insights since they show the same trends, with a slightly lower total error rate.

The relative word error rate improvements of each leave-one-speaker-out experiment using the proposed approach compared to the clean-trained baseline model are shown in Figure 5.3. For 32 out of the 35 experiments, the word error rate improved, and only for three experiments did the word error rate slightly increase. A subjective inspection did not reveal any obvious reason for the slight deterioration on these three interviews. All three interviews differ in terms of speaker style and recording situation. Most interviews are improved by 15–20 % relative to the baseline.

A more in-depth look at the 35 individual experiments is given in Figure 5.4, where we present the results of the ablation study of the proposed approach. Data augmentation has a larger impact on speech recognition than fine-tuning in many experiments. On average, the relative word error rage increase is 17.4 % removing the data augmentation from the pipeline. The data augmentation increases the word error rate for three interviews in the leave-one-speaker-out experiments.

The effects of fine-tuning are not as substantial as those of data augmentation. However, for 31 of the 35 interviews, fine-tuning further improves the recognition performance of the 3-fold model. If fine-tuning is removed from the approach, we observe an average error rate increase of 5.6 %. For five of the 35 interviews, fine-tuning results in a higher impact on the improvement than data augmentation. In

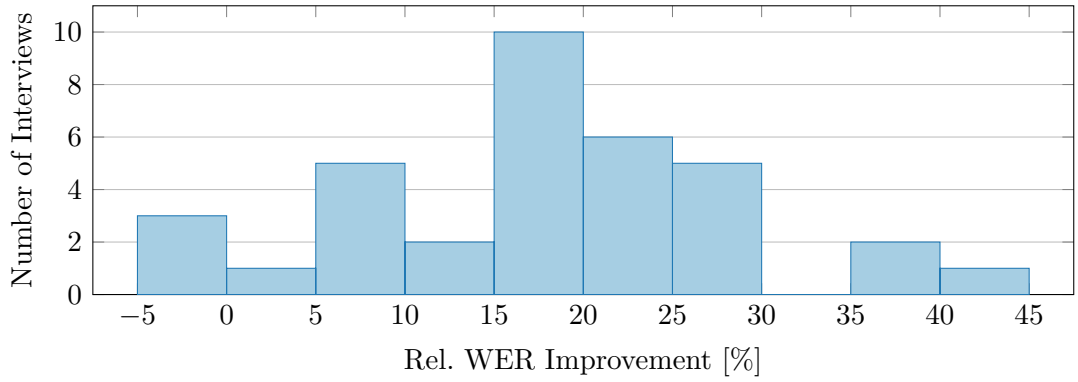


Figure 5.3: Histogram of the relative word error rate improvements with the proposed approach two-staged acoustic model adaptation compared to the clean baseline for each leave-one-speaker-out experiment with speaker-unaware decoding.

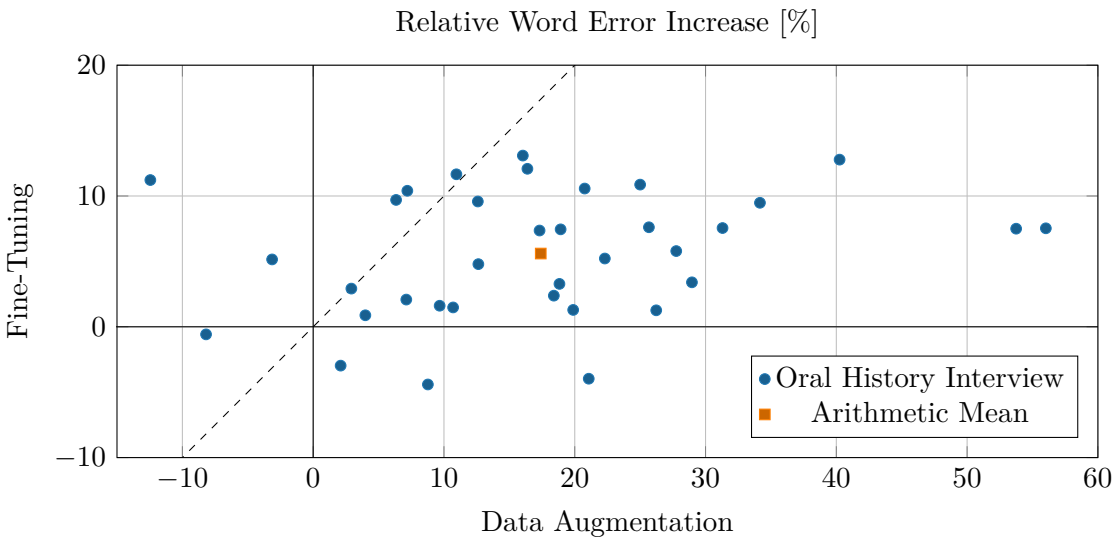


Figure 5.4: Ablation study of the two-staged acoustic model adaptation with speaker-unaware decoding by removing either data augmentation (Stage 1) or fine-tuning (Stage 2). Results are illustrated as a scatter plot of the relative word error rate increase compared to the proposed approach when one of the stages is removed. Positive values represent an increase in word error rate, i.e., the ASR performance deteriorates by removing this stage from the approach. The dashed diagonal axis marks the transition where both stages have an equal impact.

Table 5.4: Two-staged acoustic model adaptation results on test sets from different domains. Results are reported as word error rates in percent. Fine-tuned model results on the Oral History test set (marked with an asterisk) are obtained as leave-one-speaker-out cross-validation with speaker-unaware decoding, cf. Table 5.3.

Test Set	Clean	Fine-Tuned	3-fold	3-fold Fine-Tuned
GerTV Dev Set	13.8	13.6	13.7	**13.4**
DiSCo Average	12.6	12.4	**11.8**	**11.8**
Planned Clean	9.0	9.2	9.0	**8.9**
Planned Mix	11.6	11.7	**10.8**	**10.8**
Spontaneous Clean	10.3	10.1	**9.9**	**9.9**
Spontaneous Mix	19.5	18.8	17.5	**17.4**
German Broadcast 2016	12.3	11.9	11.5	**11.2**
Challenging Broadcast	20.7	20.5	20.1	**19.6**
Oral History*	33.4	31.5	28.2	**26.6**
Interaction	66.5	64.4	47.8	**47.1**
Spoken QALD-7	20.6	19.7	18.3	**17.7**

conclusion, data augmentation of the source model and fine-tuning combine well in the two-stages approach, improving speech recognition for oral history interviews.

Robustness with several evaluation sets

In this section, we investigate a possible domain overfitting by evaluating the two-staged approach on test data from other domains. We use the entire oral history set in the second stage for transfer learning for this experiment, and no data is held out. The results are shown in Table 5.4.

Even though we used the two-staged acoustic modeling adaptation to improve the performance on oral history interviews, the model performs better than the comparison models on all evaluation sets. The increase in performance is higher on rather challenging test sets while maintaining or even slightly increasing the good performance on the more clean, broadcast tasks. Therefore, we conclude that the two-staged approach is useful for adaptation to a specific task and can also help increase the generalization of the acoustic model when suitable adaptation data is used.

5.3.5 Summary and Conclusion

In this study, we presented a two-staged acoustic modeling adaptation for robust speech recognition. We evaluated the approach on our challenging German oral history interview use case. We assessed the reliability of our approach with a leave-one-speaker-out evaluation method in which we performed 35 experiments for one setup. We showed that the proposed approach increases the speech recognition performance for most experiments and performs better than using either data augmentation or fine-tuning alone. On average, the word error rate decreases relative to a clean-trained baseline by 19.3 % with speaker-aware decoding and 20.4 % with speaker-unaware decoding. Data augmentation of the source model has the most substantial influence, but fine-tuning serves as a purposeful addition.

Furthermore, we showed that our approach helps to increase the generalization of acoustic models and leads to improved recognition for challenging recordings while maintaining good performance on clean broadcast tasks. Fine-tuning in the two-step approach does not lead to domain overfitting in our experiments, despite the very small amount of training data. However, this is certainly dependent on the data used. In further experiments, we investigate the influence of the adaptation training data on speech recognition.

5.4 Study: Adaptation on Semi-Automatically Created Training Data

In this second study on transfer learning of the acoustic model, we extend the previous two-staged approach to overcome the lack of suitably annotated, *in-domain* oral history interviews for acoustic model adaptation. We utilize the robust 3-fold acoustic model to semi-automatically create training data from raw, non-segmented transcripts for the fine-tuning stage. We study two research questions in this section. First, we investigate whether fine-tuning the model, initially used to create the adaptation data, substantially improves the speech recognition for oral history interviews. Second, we study different amounts of adaptation data and the influence of the learning rate to investigate potential domain overfitting by fine-tuning.

5.4.1 Proposed Acoustic Model Fine-Tuning on Semi-Automatically Created Adaptation Data

In the previous study, we used leave-one-speaker-out cross-validation to show that the combination of data augmentation and transfer learning leads to an improved generalization of the model and, in particular, higher recognition accuracy for oral

history. However, the small amount of adaptation data in the previous experiments limits the improvement. Annotation, i.e., verbatim transcription with segmentation using timestamps every few seconds, of several hundred hours of interviews is generally required to obtain substantially improved recognition results. However, this is costly and time-intensive, which results in a lack of representative training data.

The historians at the archive *Deutsches Gedächtnis* transcribed numerous oral history interviews for their research. However, these transcripts are not directly employable for acoustic model training. The transcribed interviews are often several hours long without timestamps or temporal segmentation. Furthermore, the transcriptions are not always verbatim. They may contain formatting and comments in the text, which are not always easy to distinguish from the spoken word transcription by automatic processing. It is also not ensured that the entire interview has been transcribed or parts have been omitted.

The proposed approach to overcome the lack of German oral history ASR training data is to utilize these raw, non-segmented transcripts and semi-automatically create training data for adaptation. We extend the two-staged acoustic model adaptation idea from Section 5.3 with an automatic transcript alignment stage. The concept is illustrated in Figure 5.5. The idea is to use the robust 3-fold LF-MMI source model from Stage 1, trained on multi-conditioned broadcast speech, to perform automatic transcript alignment on the raw transcribed interviews. This data is then used as adaptation data in the subsequent stage to fine-tune the 3-fold broadcast acoustic model to the oral history domain.

In the most general sense, automatic transcript alignment is a standard component in the bootstrap training of most HMM-based speech recognition systems for short segments, cf. Section 4.3.1. However, alignment of non-verbatim, partly incomplete, hour-long speech is a substantially different task. Only certain audio parts where the transcript matches the spoken words with high confidence should be aligned. The remaining transcript and audio are to be removed. At the Fraunhofer IAIS, this type of alignment was studied by Turzynski [2017]. Automatic audio transcript alignment has also been proposed and implemented in Kaldi as a component of the *JHU Kaldi system* for the *Arabic MGB-3 ASR* challenge by Manohar et al. [2017].

We performed different primary experiments with both approaches and decided to use Manohar et al.'s approach with LF-MMI implementation for the presented research work. This automatic transcript alignment consists of segmentation and data clean-up based on the source acoustic model. The segmentation step aims at obtaining segments of a few seconds length from the hour-long interviews for acoustic model training. The data clean-up aims to remove segments or words with erroneous transcriptions.

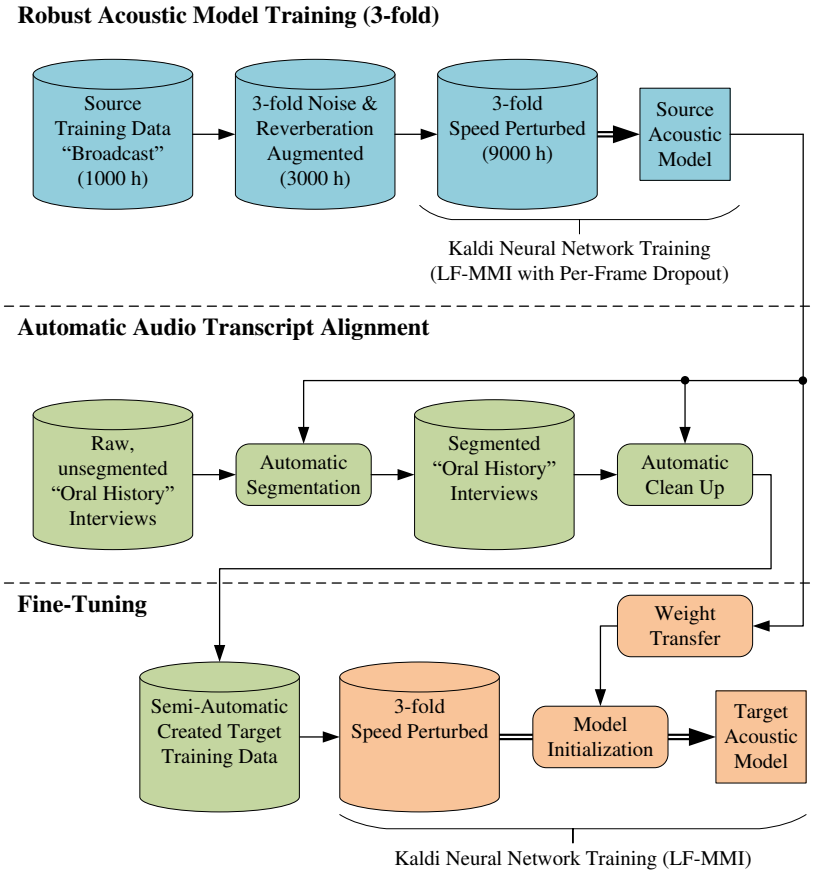


Figure 5.5: Acoustic model adaptation on automatically time-aligned speech recordings. The approach extends the two-staged adaptation using the source model to time-align and clean-up raw, transcribed but not time-aligned oral history interviews.

In their work, Manohar et al. [2017] also studied transfer learning for domain adaptation. However, their proposed approach differs from ours in substantial aspects. In our approach, we apply the automatic transcript alignment on speech that we consider out of domain for the acoustic model performing the alignment. Further, we use the semi-automatically created in-domain training data to adapt the source acoustic model to the target domain. In contrast, Manohar et al. use the automatically aligned data in their system to train a source model adapted using transfer learning on manually annotated speech.

5.4.2 Experimental Setup

Our study investigates how well the challenging oral history interviews can be processed with the automatic transcript alignment using a model trained on data-augmented broadcast speech and what improvements this adaptation data provides for the transcription of oral history interviews. Additionally, we again investigate the effect of adaptation to the oral history domain on recognition performance in other domains to estimate the real-world performance of our system. We also examine different amounts of adaptation data to determine what improvement can be expected from different amounts of adaptation data and whether a large amount of adaptation data affects performance in other domains.

In the following, we lay out our experimental setup for this study. We keep the primary experiment parameters the same as in the previous studies to make the results comparable. Therefore, the experiments are again carried out using the Kaldi ASR toolkit. We also use the same TDNN-LSTM topology with LF-MMI training we initially studied and selected in Chapter 4. Also, we use the same transfer learning setup as in Section 5.3, in particular, the same 3-fold v1 source model trained on data-augmented broadcast speech and transfer learning configuration. The decoding is performed with the default broadcast language model. For the fine-tuning experiments on 3-fold, we use the fixed language model weight as in the two-staged adaptation experiments—which is also the best configuration in most cases for oral history and the other test data. Results are reported with speaker-unaware decoding, as for most prior experiments.

The primary object of study in this experiment is the data used for automatic transcript alignment and the amount of training data for fine-tuning. We use the raw oral history interviews described in Section 3.4.9 for our experiments. The interviews were provided by the archive *Deutsches Gedächtnis*. The primary selection criterion was that the interviewees did not appear in the oral history test set. Apart from that, the selection was made randomly.

The transcripts of the interviews were provided in the original document formats. These were mainly formatted document types, such as MS Word and RTF. We performed an automatic preprocessing and text normalization to remove the

Table 5.5: Raw, transcribed but not time-aligned oral history interview data sets used for alignment and adaptation experiments. From top to bottom, the data set extends the previous one, subsequently increasing the data set size. The average interview length is reported as the arithmetic mean ± the standard deviation.

Set	Interviewees	Raw Length [hrs:min]	Raw Number of Words	Average Interview Length [hrs:min]
OH_{10}	10	44:23	291,553	4:26 ± 1:25
OH_{49}	49	195:09	1,293,103	3:58 ± 2:16
OH_{99}	99	379:09	2,451,103	3:49 ± 2:01
OH_{150}	150	546:45	3,568,885	3:38 ± 1:54

transcribers' comments from the transcription (usually, but not always in braces), write out abbreviations, numbers, and dates, remove special characters, and restore the casing of words at the beginning of sentences as good as possible. Generally, however, no manual data selection or correction of transcripts is performed due to the large amount of data. Thus, from the acoustic model training data perspective, the resulting transcriptions have to be considered partly erroneous and incomplete.

The interviews were provided and processed in four tranches. Each tranche was combined with the previous one creating a larger data set used for alignment and adaptation experiments. These four data sets are summarized in Table 5.5 with fundamental information. Due to the combination, the smaller data sets are a subset of the larger data sets, i.e., mathematically speaking, $OH_{10} \subset OH_{49} \subset OH_{99} \subset OH_{150}$, where OH_n is the data set with n interviewees. The final set comprises 150 interviews with 547 hours of raw audio recordings and more than 3.5 million transcribed words.

The average interview length is approximately 3.5 hours. However, we observe a relatively high standard deviation of almost two hours. The length of all interviews is presented in Figure 5.6 to examine the interview lengths in more detail. Overall, the interviews in the four delivered tranches seem to follow a similar underlying distribution, as indicated by the four different colors in the figure. As the amount of data increases, this underlying distribution becomes more evident. Most interviews are between 1–5 hours long, with a peak at 3–4 hours. Nine interviews are shorter than one hour, and 15 interviews, i.e., 10 % of OH_{150} are longer than 5 hours. The shortest interview recording has only 18 minutes duration. The longest, combined interview lasts more than 12 hours.

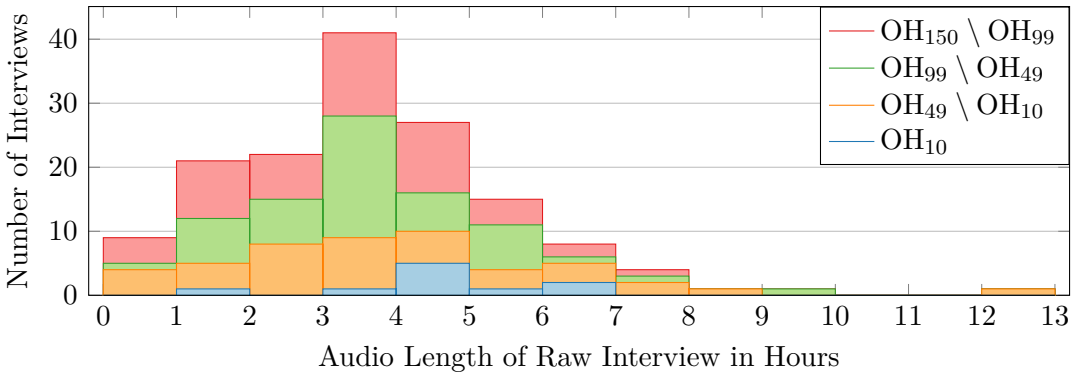


Figure 5.6: Histogram of the audio length of the raw interviews used for alignment and adaptation experiments. Each tranche is represented with a different color. Thus, the stacked bars represent the distribution of the combination of the respective tranches, i.e., the OH_n data set.

Our experiments investigate how much audio and how many transcribed words can be aligned. We cannot evaluate the alignment results objectively since we have no ground truth. However, we are anyway particularly interested in improving speech recognition. Thus, we use the four data sets for different adaptation experiments with the 3-fold model and evaluate the influence on the speech recognition performance. These experiments investigate the impact of the different data sizes and whether the automatically aligned data for fine-tuning can actually improve speech recognition.

Acoustic model training with larger training data sets is computationally expensive. Therefore, an exhausting hyperparameter optimization for all adaptation sets is not feasible within the scope of this work. However, to investigate the influence of the learning rate on acoustic model adaptation and, in particular, the impact on different domains, we perform representative experiments on the medium-sized, transcript-aligned oral history adaptation set OH_{49}.

5.4.3 Results and Discussion

In the following, we present the results of our experiments. We first present the results of the automatic transcript alignment. Then, we present the speech recognition results when using the aligned data for fine-tuning.

Automatic Transcript Alignment

The result summary of the automatic transcript alignment for the four interview sets is presented in Table 5.6. Depending on the data set size, 39–47 % of the overall audio was aligned using the 3-fold LF-MMI model. However, the number of aligned

Table 5.6: Results of the automatic transcript alignment of the oral history interviews. Results are reported for the four data sets. The 3-fold LF-MMI model was utilized for the alignment.

	$\mathbf{OH_{10}}$	$\mathbf{OH_{49}}$	$\mathbf{OH_{99}}$	$\mathbf{OH_{150}}$
Overall Aligned Audio	47.4 %	38.9 %	42.4 %	45.5 %
Overall Aligned Words	67.9 %	55.4 %	60.7 %	64.4 %
Length [hrs:min]	21:02	75:58	160:55	248:34
Number of Segments	18,231	64,704	138,532	213,783
Number of Words	198,059	715,993	1,488,271	2,297,880
Unique Words	15,666	36,935	60,026	80,421
Average Segment Length [s]	4.2 ± 2.4	4.2 ± 2.3	4.2 ± 2.4	4.2 ± 2.4
Average Words Per Segment	10.9 ± 7.3	11.1 ± 7.0	10.7 ± 7.1	10.7 ± 7.2
Average Words Per Second	2.6 ± 0.9	2.6 ± 0.9	2.5 ± 0.9	2.5 ± 0.9

words is in the range of 55–68 %, thus, substantially larger than the aligned audio. This is because many interviews contain passages in which interviewees do not speak. The alignment removes these audio passages.

Overall, slightly more than half of the spoken, transcribed words were aligned with the audio and can be used for acoustic model training. The smallest set with ten interviews comprises about 21 hours of training data, the largest set with all 150 interviews comprises almost 250 hours. The overall length of the data set and the number of words is about one-fourth of the GerTV1000h corpus, cf. Table 3.2. The vocabulary size (number of unique words) is about one-third of the GerTV1000h corpus. The four oral history adaptation sets are quite similar in terms of statistical properties. They also have similar properties to the broadcast GerTV1000h training data, cf. Table 3.4. The average segment length, words per segment, and words per second are slightly lower for the aligned oral history interviews than for GerTV1000h.

Figure 5.7 shows the ratio of aligned audio to the raw audio length per interview in more detail. We observe a maximum in interviews whose audio could be aligned to 50 to 60 %. Overall, however, the distribution is quite wide, and most interviews fall into the range of 30–80 %. Only for some interviews, more than 80 % of the audio could be aligned.

However, some interviews could practically not be aligned at all. A qualitative evaluation of these interviews showed that the failed alignments are partly due to incomplete or non-matching transcripts and extremely difficult, oral-history-specific characteristics. These are mainly dialects and challenging recording con-

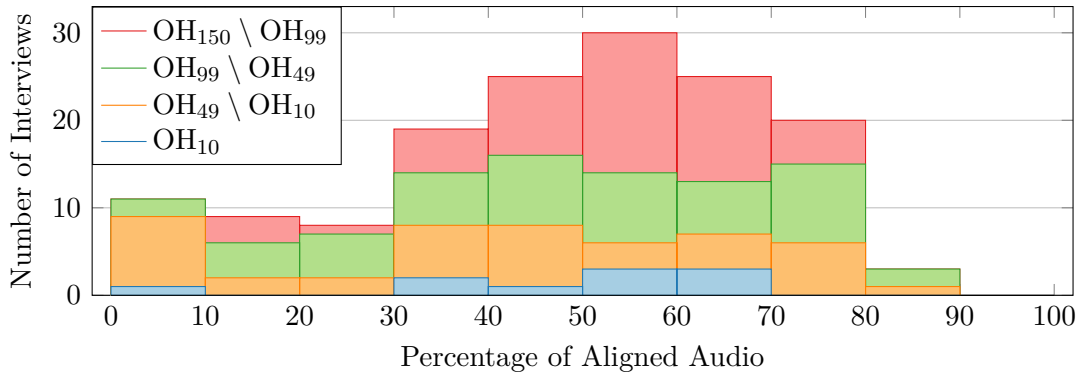


Figure 5.7: Histogram of the aligned audio length of the raw interviews. Each tranche is represented with a different color. Thus, the stacked bars represent the distribution of the combination of the respective tranches, i.e., the OH_n data set.

ditions. Some are barely intelligible even for humans. Although the data is initially lost for model adaptation, the reliable discarding of the non-alignable interviews indicates the general reliability of the automatic transcript alignment approach even for challenging oral history interviews. Furthermore, it indicates the importance of robustness of the source model required for the alignment, as acoustic challenges often seemed to be the primary reason for the failed alignment.

Acoustic Model Fine-Tuning Results

The results of the adaptation experiments with the four semi-automatically created oral history training data sets are reported in Table 5.7. We present the results compared to two baselines: the source 3-fold model and a comparable LF-MMI–TDNN-LSTM trained from scratch on the largest oral history data set. As presented in Table B.2 in the appendix, we also evaluated training from scratch on the smaller subsets for comparison. However, for the sake of clarity, we only report the baseline with the best results using the largest data set. We also again report the results of the adaptation experiments on the oral history test set from Section 5.3 for comparison. We refer to this data as T_{35}.

First, it is noteworthy that the baseline system trained from scratch on 250 hours of transcript-aligned oral history interviews already achieves a word error rate on the oral history test set comparable to the 3-fold model trained with substantially more data. This indicates an overall decent quality of the transcript alignment that enabled this recognition performance.

The results show that adaptation on data created with automatic transcript alignment generally improves recognition performance on the oral history test set. The more data used, the higher the improvement. There is also a slight

Table 5.7: Results of acoustic model adaptation experiments with oral history training data created with automatic transcript alignment. Results are reported as word error rates in percent. T_{35} refers to the adaptation using the oral history test set presented in Section 5.3. The result of this model on the oral history test set, marked with an asterisk, was obtained using the leave-one-speaker-out cross-validation approach.

	Baselines		**3-fold v1 Fine-Tuned**				
	OH_{150}	**3-fold v1**	**T_{35}**	**OH_{10}**	**OH_{49}**	**OH_{99}**	**OH_{150}**
Training Data Size:	249 h	3×992 h	3.5 h	21 h	76 h	161 h	249 h
GerTV Dev Set	17.9	13.7	**13.4**	13.6	13.8	13.8	13.8
DiSCo Average	21.0	**11.8**	**11.8**	12.3	12.7	12.7	12.7
Planned Clean	15.1	9.0	**8.9**	9.2	9.5	9.6	9.5
Planned Mix	21.3	**10.8**	**10.8**	11.3	11.4	11.4	11.4
Spontaneous Clean	16.2	**9.9**	**9.9**	10.5	11.4	11.6	11.4
Spontaneous Mix	31.4	17.5	**17.4**	18.4	18.4	18.4	18.4
German Broadcast 2016	15.9	11.5	**11.2**	11.8	11.3	**11.2**	11.3
Challenging Broadcast	26.3	20.1	19.6	19.8	**19.5**	**19.5**	**19.5**
Oral History	28.6	28.2	*26.6	25.7	25.2	24.7	**24.5**
Interaction	58.6	47.8	**47.1**	47.7	48.6	49.2	49.1
Spoken QALD-7	28.4	18.3	17.7	17.7	18.1	**17.5**	17.8

cross-domain improvement with the semi-automatically generated training data on Challenging Broadcast and Spoken QALD-7. However, on the DiSCo subsets and Interaction, the model adapted with T_{35} continues to be the best performing system. This might indicate some domain overfitting with an increased adaptation data set size.

The improvements relative to the results with the 3-fold source model per test set are compared in Figure 5.8. In this plot, the effect of adaptation and the influence of training data size becomes evident. We observe a monotonically increasing but flattening curve with increasing adaptation data size for oral history. The relative improvement with 76 hours of data is already above 10 % and becomes about 12–13 % with 161 and 249 hours. Except for small outliers, we observe mainly consistent behavior and slight improvements around 2.5 % relative for Broadcast 2016, Challenging Broadcast, and Spoken QALD-7. DiSCo and Interaction deteriorate almost monotonically with increasing adaptation data. However, the curve for both seems to saturate at 161 hours. Further experiments are required to reasonably estimate the behavior with more in-domain adaptation data, which is currently unavailable.

We want to investigate further the effects of adaptation on the different domains. We are particularly interested in why the recognition on the Interaction test sets

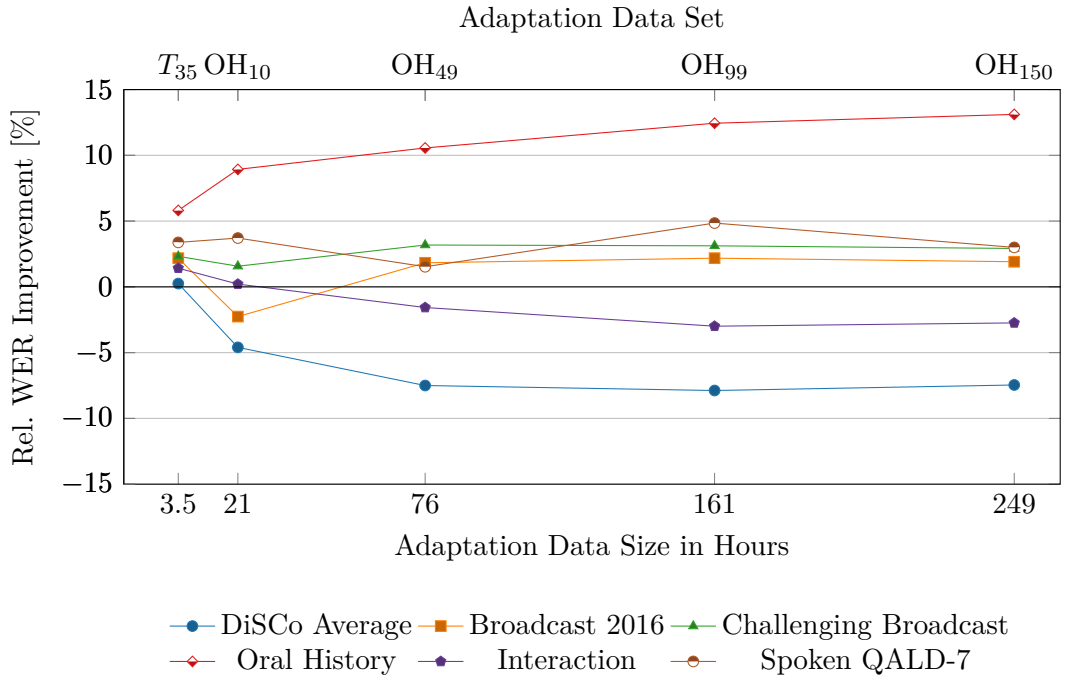


Figure 5.8: Relative word error rate improvements of the acoustic models adapted with increasing oral history adaptation data size on different test sets. All values above zero indicate an improvement in the recognition performance. The result on Oral History with T_{35} was obtained using a leave-one-speaker-out evaluation in Section 5.3.

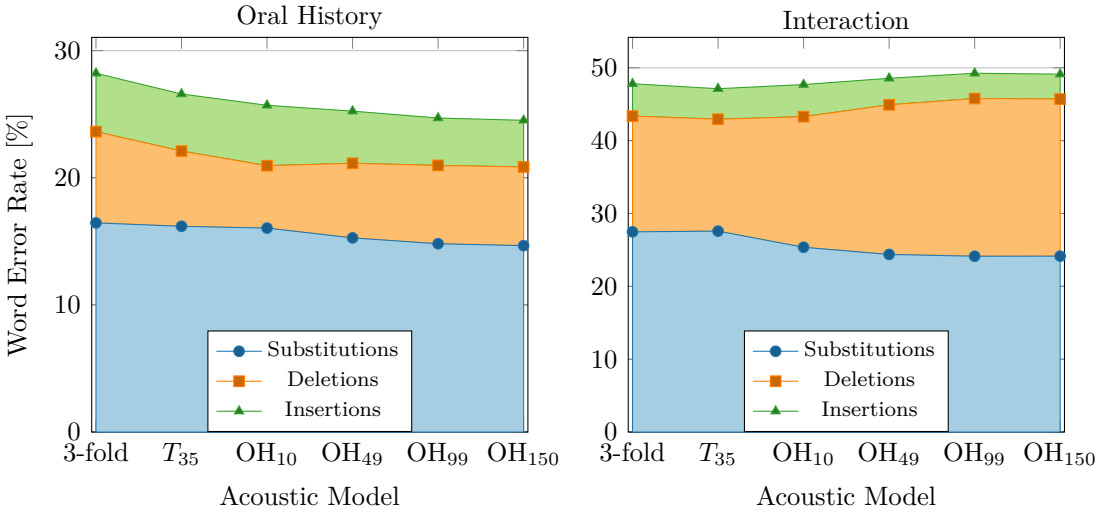


Figure 5.9: Influence of word error types on the word error rate for Oral History and Interaction with different adapted models. The different error types are stacked in the graph so that the stacked graph represents the total word error rate. Each color represents the influence of the respective error type on the total word error rate.

deteriorates with adaptation, although Interaction and Oral History share many similar challenges. Figure 5.9 shows the respective influence of substitutions, insertions, and deletions on the word error rate for the Oral History and Interaction test set for each model.

Substitutions account for the largest impact in each case for both test sets. This is typical for challenging speech recognition test data with unconstrained vocabulary. Adaptation with the oral history training data reduces the influence of substitutions for both test data so that words are recognized more precisely.

Insertions have only a slight influence on both test sets. Strikingly, deletions have a comparatively substantial impact on the overall word error rate for Interaction. In particular, short words are often not recognized. This issue becomes more substantial for the Interaction test set with adaptation towards the oral history domain, while it has little influence on the Oral History test set. In particular, short words are recognized more poorly for Interaction with increasing adaptation. As summarized in Section 3.4.11, Oral History and Interaction differ, particularly in the speed of speech. While Oral History has one of the slowest average speaking rates, Interaction has the highest speaking rate in our data collection. We observed that a high speaking rate increases deletions in our ASR systems. The adaptation with oral history seems to contribute to the systems' accuracy decreases with increasing speech tempos.

Table 5.8: Results of experiments with different learning rate configurations for fine-tuning with transcript-aligned oral history interviews. The adaptation experiments were carried out with the oral history training set OH_{49}. Results are reported as word error rates in percent compared to the 3-fold source model. X/Y refers to the initial and final learning rate used for fine-tuning. 1e-6/1e-7 is the default configuration.

		Fine-Tuned (LR)		
Test Set	**3-fold**	**1e-7/1e-8**	**1e-6/1e-7**	**1e-5/1e-6**
GerTV Dev Set	**13.7**	**13.7**	13.8	14.2
DiSCo Average	**11.8**	12.0	12.7	13.5
Planned Clean	9.0	**8.9**	9.5	10.1
Planned Mix	**10.8**	10.9	11.4	12.1
Spontaneous Clean	**9.9**	10.4	11.4	11.9
Spontaneous Mix	**17.5**	17.7	18.4	20.1
German Broadcast 2016	11.5	**11.3**	**11.3**	11.6
Challenging Broadcast	20.1	**19.5**	**19.5**	20.4
Oral History	28.2	26.3	**25.2**	25.5
Interaction	47.8	**47.1**	48.6	49.8
Spoken QALD-7	18.3	**17.3**	18.1	21.3

Influence of the Learning Rate

We used the same learning rate configuration for all experiments, initially established in preliminary experiments with small data sets. The adaptation set size was increased by ten-fold or more in the current experiments. On the one hand, the question arises whether an increase in the amount of training data is essential for the observed improvements or whether an increase in the learning rate with a fixed amount of training data leads to similar improvements. On the other hand, the question arises whether a lower learning rate is more effective in avoiding domain overfitting when the adaptation data is increased multi-fold. To investigate the influence of re-adjusting the learning rate, we have examined different learning rates with the medium-sized data set OH_{49}. The results are presented in Table 5.8.

The results show that the best results on oral history are achieved with our proposed default learning rate of 1e-6/1e-7. Increasing the learning rate to 1e-5/1e-6 leads to slightly worse results on oral history. However, it substantially deteriorates all other test sets and domains—almost always worse than the 3-fold

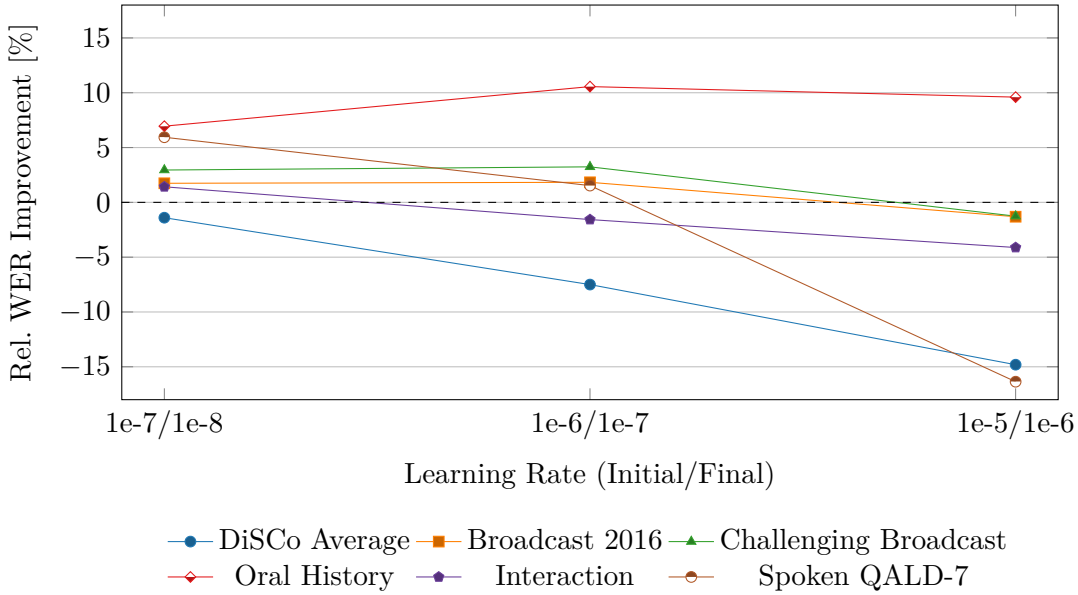


Figure 5.10: Relative word error rate improvements of fine-tuning with different learning rates on the semi-automatically created oral history training set. The experiments were carried out with the OH_{49} set.

baseline. Since this is true across domains, except for oral history, this indicates a significant domain mismatch due to fine-tuning towards oral history.

Reducing the learning rate to 1e-7/1e-8 leads to better results on the other-domain test sets Interaction and Spoken QALD-7 than the default learning rate and the 3-fold baseline. An improvement over the 3-fold baseline is also observed for oral history and the broadcast domain. A reduced learning rate with the oral history OH_{49} adaptations set thus seems to generally achieve a slight but consistent improvement and a further reduction in domain mismatch over the 3-fold model.

As shown in Figure Figure 5.10, the influence of the different learning rates on different domains becomes even more evident when considering the improvement per test set relative to the source model. With the lowest learning rates, the relative improvements are almost all in the range from −2 to 5 %, with the biggest relative improvement on Oral History and deterioration only on DiSCo. As the learning rate increases, the domain mismatch seems to increase. With the highest learning rate, only oral history improves. Overall, the proposed learning rate of 1e-6/1e-7 seems to be an appropriate compromise to achieve adequate improvements in the oral history domain without risking a substantial domain mismatch.

Compared to Figure 5.8, where we studied different data set sizes with a fixed learning rate, increasing the training data set seems more suitable to avoid domain mismatch than increasing the learning rate. Since only three learning rates were

investigated in our study, we can assume a better learning rate exists between the investigated ones. However, training acoustic models is computationally and time-intensive. An exhaustive hyperparameter optimization for all learning rates, especially individually for each data set size, is not feasible within the scope of this work. In Section 5.5.2 we continue investigating the influence of the learning rate in additional adaptation experiments with different training data sizes.

5.4.4 Summary and Conclusion

In summary, our experiments in this study showed that automatic transcript alignment is well suited to overcome the lack of training data and to adapt an out-of-domain acoustic model to a target domain in the same language. A substantial amount of hours-long, raw transcribed but not time-aligned oral history interviews were time-aligned to create in-domain training data semi-automatically. These interviews were used to fine-tune the robust 3-fold broadcast acoustic model.

Our contribution in this study is to investigate automatic transcript alignment and adaptation for out-of-domain data in the same language. We further studied the influence of the adaptation data size and the learning rate on the overall performance and domain overfitting using several test sets from different domains. Overall, the adapted models generalize well and are thus also able to better handle similar challenges in other domains.

Our experiments show that the recognition performance on oral history improves monotonically with larger semi-automatically generated training data sets. Only for Interaction and DiSCo, which is closest to the broadcast training data, we observe a deterioration of the recognition performance due to the proposed adaptation. We observe a fairly consistent improvement for our other test sets from different domains with different oral history adaptation data sets. Overall, we substantially improved the recognition on the Oral History test set to a 24.5 % word error rate with the best adaptation.

5.5 Study: Domain-Mismatch within the Oral History Domain

Towards the end of the present research work in 2021, an additional oral history corpus for automatic speech recognition became available. These annotated interviews were provided by the *Haus der Geschichte* (House of the History) Foundation (HdG) in a joint, interdisciplinary research project. The HdG data sets differ substantially from the oral history interviews provided by the *Deutsches Gedächtnis* archive in the KA^3 project, primarily studied in the presented research work. As discussed in Section 3.4.10, the HdG interviews were recorded more recently and

thus have substantially better audio quality. Therefore, the HdG data cover a specific subdomain in the oral history domain and can be considered a *Clean Oral History* data set. In contrast, the primarily considered (transcript aligned) KA^3 training and test interviews represent *Mixed Oral History* data.

This new data allows us to investigate interviews in the oral history domain more nuancedly in the following study. In particular, we investigate the influence of the acoustic recording conditions of oral history interviews on the acoustic model and domain adaptation. The experiments in this study are primarily based on [Gref et al., 2022b].

5.5.1 Experimental Setup

For the experiments in this section, we utilize the latest robust 3-fold acoustic model, 3-fold v2, as the source model and baseline. Thus, the results of these experiments are not directly comparable to the previous experiments in the presented research work, as the 3-fold v2 model generally performs better on all domains than the primary used 3-fold v1 model, cf. Section 4.4.5. Furthermore, for evaluation, we do not use the default language model but the *large* 5-gram general-purpose broadcast language model, which was the standard language model at that time, cf. Section 3.5.

We use the 6:20-hour speaker-independent training split of the HdG corpus representing the Clean Oral History domain for adaptation. As Section 3.6 describes, we obtained three annotations of three different transcribers for each segment in the HdG corpus. Since there is no reasonable way to automatically merge the three different transcriptions into one for training without manual post-annotation, we use the transcription of transcriber A for training. Since this might lead to a bias in the evaluation towards the annotation style of A, the evaluation with HdG data is always done on the transcriptions of all transcribers A, B, and C. In this way, bias or overfitting towards transcriber A can be identified and prevented.

Additionally, we use the semi-automatically created OH_{150} data set from Section 5.4 with 250 hours of transcript-aligned interviews from 150 different speakers for adaptation. This data set represents the Mixed Oral History domain. Both data sets differ substantially in size. Therefore, we additionally study adaptation with a 10 % subset of OH_{150}. By comparison of the 250-hour and this 25-hour subset $OH_{150}^{10\%}$, we investigate the effect of a possible domain overfitting to the acoustic conditions of the Mixed Oral History domain when training on 250 hours. We further investigate the combination of this mixed data set with the comparatively clean HdG training data set.

Except for the language model, the source 3-fold model, and the adaptation data, we use the precisely same setup as in Section 5.4 for these experiments. In particular, we use our proposed default learning rate, which is a reasonable

Table 5.9: Comparison of acoustic model adaptation experiments using different oral history adaptation and test sets. Results are reported as word error rate in percentage. HdG Test and Dev. Average (Avg.) are the respective arithmetic mean ± the standard deviation of the results of the ASR system on the three different human transcriptions. Additionally, we also report the respective results for Transcriber A, B, and C as the reference.

		Adaptation Data			
Test Set	**3-fold v2 Baseline**	**HdG Tr. A**	**$OH_{150}^{10\%}$**	**HdG +$OH_{150}^{10\%}$**	**OH_{150}**
(Mixed) Oral History	26.0	25.7	24.7	24.6	**23.9**
HdG Dev. Avg.	17.3±1.06	17.0±1.08	16.7±1.02	**16.6**±1.08	17.1±1.09
Transcriber A	16.7	16.4	16.2	**16.0**	16.4
Transcriber B	16.6	16.3	16.1	**16.0**	16.5
Transcriber C	18.5	18.2	**17.9**	**17.9**	18.4
HdG Test Avg.	16.4±0.32	15.9±0.30	**15.6**±0.33	15.7±0.37	16.1±0.36
Transcriber A	16.1	15.6	**15.3**	**15.3**	15.8
Transcriber B	16.4	15.9	**15.6**	15.7	16.1
Transcriber C	16.8	16.2	**16.0**	16.1	16.5

compromise for adaptation in terms of domain overfitting in Section 5.4.3. In this study, we again discuss an adjustment of the learning rate. In detail, we investigate a reduction of the learning rate as an alternative approach to reducing the size of the adaptation data set OH_{150} to $OH_{150}^{10\%}$.

5.5.2 Results and Discussion

In the following, we present and discuss the results of the adaptation experiments. We first present the results on the different oral history test sets. Then we present the results of the adapted models on the test sets from other domains to estimate the real-world performance of the models for unseen data. We conclude with a discussion of the learning rate adjustment.

Results on Oral History Test Sets

All adaptation experiments on the different oral history test sets are summarized in Table 5.9. The experiments show that adaptation of the acoustic model can lead to consistent speech recognition improvements in the overall oral history domain.

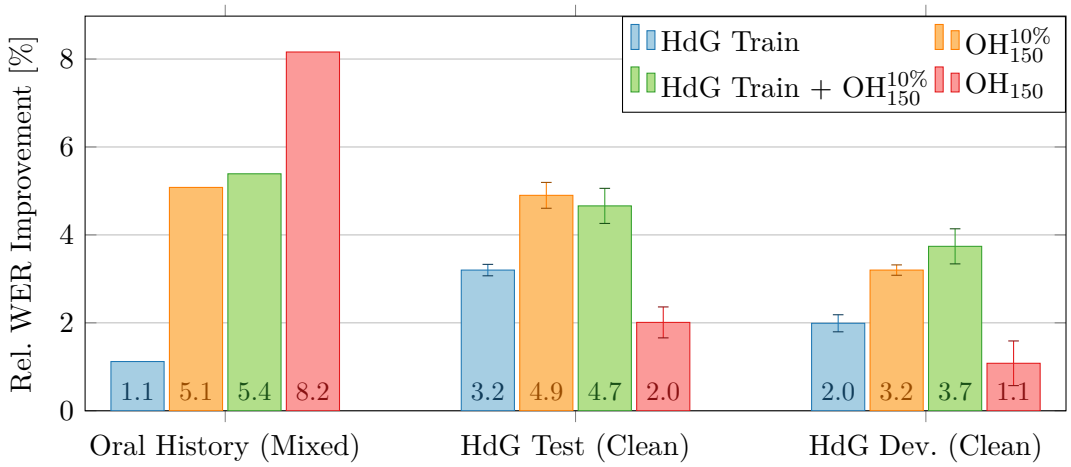


Figure 5.11: Relative improvements of the four differently adapted models on the three different oral history test sets. The error bars for the HdG sets represent the standard deviation of the three different reference transcripts used for evaluation.

However, the overall improvement is dependent on the adaptation data. In particular, more data does not necessarily lead to better recognition performance. With the 250 hours of forced aligned data for adaptation, we achieve an improvement of 8.2 % relative to the 3-fold v2 broadcast base model on Mixed Oral History Test—our conventional, primary used test set in this work. This adapted model outperforms our prior best model from Section 5.4, adapted on the same OH_{150} data, by 1.6 percentage points, due to the improved 3-fold v2 source acoustic model and larger language model.

An overview of the relative improvements of the models is shown in Figure 5.11. All models improve the recognition performance on all test data. However, adaptation on the largest 250-hour data set OH_{150} has wildly different effects on the two oral history domains. While it substantially improves recognition performance for interviews with challenging acoustic conditions, it achieves the least improvement on the clean HdG oral history domain. On the HdG development and test set, the relative improvement of this model is in the range of 1–2 %.

In contrast, adaptation on the HdG training split of Transcriber A results in a 2.0 % relative improvement on HdG Dev and 3.2 % on Test. This is remarkable, as only 6.35 hours instead of 250 hours of training data is used. The adaptation improvements are very similar for Transcriber A, B, and C. Thus, the improvement on the HdG data is consistent and not just a bias towards the transcription style of A. However, the improvement of this adaptation on the mixed test set is only 1.1 % relative to the baseline.

The greatest improvement on the clean HdG test interviews is obtained utilizing the 25-hour subset $\mathrm{OH}_{150}^{10\%}$ with interviews in mixed acoustic conditions. Although this data set has substantially worse acoustic recording conditions and contains semi-automatic transcriptions, the acoustic model seems to generalize better with this data. For HdG Dev, the relative improvement is 3.2 % relative to baseline, and for HdG Test 4.9 %. On Mixed Oral History, we achieved a relative improvement in a similar range of 5.1 %. Therefore, we infer that this training data set size with interviews in mixed acoustic conditions seems a reasonable compromise to cover both clean and mixed interviews by adapting the acoustic model.

Combining the 25-hour training data set with the clean HdG interviews slightly improves further the performance on Mixed Oral History, and the HdG Dev set. However, on HdG Test, the performance is somewhat decreased. Overall, the differences in recognition performance by adding the HdG training set to $\mathrm{OH}_{150}^{10\%}$ are not very substantial. Although adding the HdG interviews in this setup slightly increases the variance for the different annotators, overall, the results remain consistent for all reference transcribers.

Overall, we conclude that substantial improvement on oral history interviews can be achieved with comparatively few hours of adaptation data, both semi-automatically and manually annotated. Furthermore, we infer a large data set can lead to overfitting to the recording conditions in the domain, as in the case of the mixed oral history interviews. Thus, depending on the application and the type of data, it may be valuable to experiment with varying subsets of data. Overall, a relative improvement of about 5 % can be achieved by adapting the acoustic model for each of the two different oral history test data.

Compared to the human word error rate of 8.7 percent for the HdG data that we worked out in Section 3.6, speech recognition still has quite a way to go to achieve human performance on oral history data. The error rate has to be roughly halved until an ASR system can replace manual transcription in most scenarios and make human transcriptions superfluous. However, the current recognition performance of the systems is already sufficient so that after a manual correction, the transcript can be used for the *Zeitzeugenportal* of the *Haus der Geschichte* Foundation. The transcripts are essential documents for the practical use of oral history interviews. They are primarily used to index the videos' contents for the thematic classification on the online service *Zeitzeugenportal*. Since oral history videos are also a meaningful component of the exhibition practice in museums, the transcripts are also used for cut lists. Additionally, the transcripts serve for the subtitling of the videos.

Furthermore, our analysis of two different oral history corpora uncovers a substantial difference in speech recognition performance. The Mixed Oral History test data is much more challenging for the ASR system than the HdG data—even

Table 5.10: Comparison of acoustic model adaptation experiments on test sets from other domains. Results are reported as word error rate in percentage.

Test Set	3-fold v2 Baseline	Adaptation Data			
		HdG Tr. A	$\mathrm{OH}_{150}^{10\%}$	HdG +$\mathrm{OH}_{150}^{10\%}$	OH_{150}
GerTV Dev Set	**12.8**	**12.8**	**12.8**	**12.8**	13.1
DiSCo Average	**12.1**	12.2	12.4	12.5	12.9
Planned Clean	**8.8**	**8.8**	8.9	9.0	9.2
Planned Mix	**9.8**	9.9	10.0	10.0	10.4
Spontaneous Clean	**11.1**	11.3	11.5	11.6	11.9
Spontaneous Mix	**18.8**	19.0	19.3	19.4	20.0
German Broadcast 2016	9.2	**9.1**	**9.1**	**9.1**	9.5
Challenging Broadcast	**17.2**	17.3	17.4	17.5	18.0
Interaction	**49.6**	50.1	50.3	50.2	51.3
Spoken QALD-7	13.6	13.2	13.3	13.0	**12.4**

when adapted with 250 hours of additional representative data from the very same source. Depending on the model, the absolute difference in the word error rate between the mixed and clean interviews is 9.5 % (3-fold v2), 7.7 % (OH_{150}), and 8.9–9.0 % ($\mathrm{OH}_{150}^{10\%}$ with/without HdG train). Both data sets from the German oral history domain have similar characteristics of speakers, especially in terms of age, language, dialects, and topics. The main difference lies in the wide range of recording age of the mixed interviews and the resulting acoustic recording quality. Therefore, for our models, approximately 9 % of word error rate percentage points are still attributable to the acoustic challenges of oral history. Thus, further improving acoustic robustness for oral history remains an open field of research, although substantial improvements have been made.

Results on Test Sets from Other Domains

In Table 5.10, we summarize the results of the adapted models on the test sets from domains other than oral history. Similar to the adaptation experiments in Section 5.4, the adaptation towards the oral history domain decreases the recognition performance on DiSCo and Interaction. However, in this experiment, we also observe a slight deterioration instead of improvement in recognition performance on Challenging Broadcast, cf. Table 5.7. In particular, this deterioration is strongest for adaptation with OH_{150}. This is noteworthy as the adaptation of the

3-fold v1 source model with this data improved the performance on this test set. However, the recognition accuracy of the source model 3-fold v2 with the large language model used in this experiment is already substantially better than the 3-fold v1 baseline, cf. Table 4.14. Therefore, the observed behavior is probably because the improved 3-fold v2 source model already models the acoustic challenges of Challenining Broadcast quite well. Further, non-domain-specific adaptation seems detrimental to the robustness in this domain.

On the Spoken QALD-7 test set, we achieve a relative improvement similar to the oral history test sets with the different adaptations. With the 250-hour OH_{150} adaptation, a relative improvement of 9.0 % is achieved. This is a new benchmark on this test set. With this model, the word error rate is now roughly in the range of DiSCo Average and GerTV Dev. We consider this a successful reduction of the domain mismatch between broadcast and the speech assistant domain.

Discussion on Learning Rate Reduction vs. Data Set Reduction

In the previous evaluation, the 10 % subset $\mathrm{OH}_{150}^{10\%}$ of the 250-hour oral history training corpus OH_{150} proved a reasonable compromise for adaptation for the different oral history domains. Adaptation with 250 hours resulted in domain overfitting for interviews in mixed acoustic conditions. However, we used a fixed learning rate for all experiments. Instead of reducing the data set size to 10 %, one could also reduce the learning rate accordingly to avoid domain overfitting. This is investigated in the following experiment.

As a comparison to the 25-hour subset, we train a model with the entire 250 hours of data and a learning rate reduced by the same factor, i.e., 1e-7/1e-8 instead of the default 1e-6/1e-7 learning rate setup. The results are shown in Table 5.11 compared to the models trained on $\mathrm{OH}_{150}^{10\%}$ and OH_{150} with the default learning from the previous section.

As to be expected, the 250-hour model with ten times reduced learning rate has almost the same word error rate for the test sets from the broadcast domain and as the 25-hour model with default learning rate. Likewise, it is narrowly better for Interaction, Spoken QALD-7, and (Mixed) Oral History. As more and different utterances are used for training, a slightly better generalization of this model with more data and the corresponding adjusted learning rate is to be expected. Remarkably, however, we see a substantial difference on the HdG interviews in favor of the model trained on $\mathrm{OH}_{150}^{10\%}$ with less training data and the default learning rate.

For the different oral history evaluation and development sets, the relative improvements compared to the 3-fold v2 source model are illustrated in Figure 5.12. On (Mixed) Oral History, the relative improvement of $\mathrm{OH}_{150}^{10\%}$ with the default learning rate and OH_{150} with the reduced learning rate are very similar. However,

Table 5.11: Comparison of adaptation experiments with reduced learning rate instead of reduced training data size. 1e-6/1e-7 is the default learning rate setup, 1e-7/1e-8 is the reduced learning rate. The detailed results for each reference annotator on the HdG sets are reported in Table B.3 in the appendix.

Adaptation Data Set	$OH_{150}^{10\%}$	OH_{150}	OH_{150}
Adaptation Data Size	25 h	250 h	250 h
Learning Rate	1e-6/1e-7	1e-7/1e-8	1e-6/1e-7
GerTV Dev Set	**12.8**	12.9	13.1
DiSCo Average	**12.4**	**12.4**	12.9
Planned Clean	**8.9**	9.0	9.2
Planned Mix	10.0	**9.9**	10.4
Spontaneous Clean	**11.5**	**11.5**	11.9
Spontaneous Mix	19.3	**19.2**	20.0
German Broadcast 2016	**9.1**	**9.1**	9.5
Challenging Broadcast	**17.4**	**17.4**	18.0
(Mixed) Oral History	24.7	24.6	**23.9**
HdG Dev. Avg.	**16.7**±1.02	17.0±1.07	17.1±1.09
HdG Test Avg.	**15.6**±0.33	15.9±0.32	16.1±0.36
Interaction	50.3	**50.2**	51.3
Spoken QALD-7	13.3	13.2	**12.4**

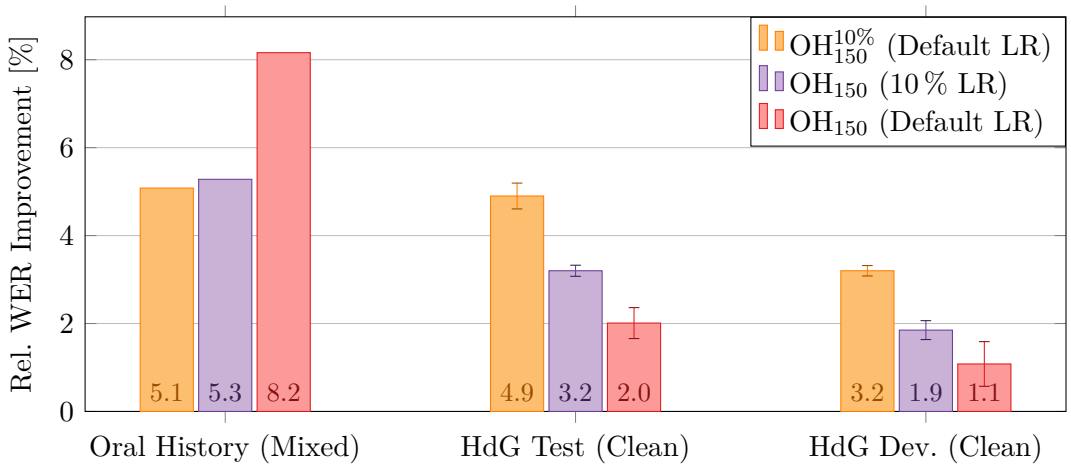


Figure 5.12: Comparison of the relative improvements of the learning rate discussion on the three different oral history test sets. The error bars for the HdG sets represent the standard deviation of the three different reference transcripts used for evaluation. The same colors as in Figure 5.11 are used for the reference models with the default learning rate.

for both HdG sets, the results of OH_{150} with the reduced learning rate lies between $OH_{150}^{10\%}$ and OH_{150} with the default learning rate. Overall, the absolute adaptation data set size seems less crucial for domain adaptation than an appropriate learning rate choice. Using a subset achieves similarly good or better results and simultaneously saves training time.

5.5.3 Summary and Conclusions

The presented study investigates the influence of oral history training data from different domains on the acoustic model adaptation via fine-tuning. We utilized a new, *clean* oral history corpus provided in 2021 that primarily contains oral history interviews with high recording quality. We utilized this corpus for adaptation and evaluation, studying the influence of acoustic recording conditions of oral history interviews on the proposed adaptation and the overall performance of the speech recognition systems.

In particular, we showed that even with just 25 hours of adaptation data, a consistent improvement by 5 % relative to a robust baseline was achieved for the different oral history domains. Utilizing 250 hours of adaptation data in diverse, mixed acoustic recording conditions leads to more substantial improvements on similar data—but only to minimal improvements on clean oral history interviews. Thus, a large amount of adaptation data might not necessarily lead to good generalization but instead might lead to domain overfitting. On the other hand, we achieved relative improvements in the range of 2.0–3.2 % in the same domain with as little as six hours of adaptation data for clean oral history interviews.

By comparing the results on the clean and mixed oral history test data, we inferred that, on average, roughly nine percentage points of the word error rate of our systems are attributed to challenging acoustic recording conditions. Although we achieved substantial improvements in acoustic robustness in the presented research work, especially in Chapter 4, this issue is not yet fully resolved. It remains one of the core challenges for oral history interviews.

Furthermore, we have shown that adaptation of the acoustic model toward the oral history domain also influences the performance of speech data from other domains. Adaptation leads to degradation in the broadcast domain for a robust source model that works very well on this domain. However, the adaptation simultaneously leads to substantial improvements in the speech assistant domain, which also has acoustic challenges.

In conclusion, it appears that for acoustic model adaptation via fine-tuning, *quality over quantity* applies, whereby quality does refer to the audio quality but rather how accurately the adaptation data matches the target domain. Naturally, with only a few hours of additional training data, no enormous improvement of the ASR can be achieved. However, the proposed adaptation allows improvement with

comparatively little transcription effort. With human correction of automatically generated transcripts, such an improvement can save many person-hours when processing vast amounts of data, as is common in many oral history archives. These can further improve the speech recognition system in an iterative adaptation process.

5.6 Summary and Contributions

5.6.1 Summary

In this chapter, we investigated fine-tuning of the LF-MMI trained DNN-HMM acoustic model as an additional method for acoustic robustness and domain adaptation in three different studies. We first studied the method as a two-staged acoustic model adaptation approach using a leave-one-speaker-out evaluation with the 3.5-hour oral history data set. With this approach, we improved the average word error rate on this set of the 3-fold v1 model from 28.2 % in the last chapter to 26.6 %, simultaneously improving recognition performance on speech recordings from other domains. The small amount of adaptation data primarily limited the improvement.

We studied automatic transcript alignment to semi-automatically generate adaptation data from raw transcribed but not time-aligned oral history interviews. Despite the challenging recording conditions, we have demonstrated that suitable adaptation data can be generated from the raw transcribed interviews without further manual correction. We further improved the word error rate on the oral history test set to 24.5 % with the best adaptation of the 3-fold v1 model. With the adaptation of the 3-fold v2 model, we achieved a word error rate of 23.9 %.

Additionally, we investigated the influence of the amount and type of adaptation data and the learning rate on improving the acoustic model for target domain and domain overfitting. Our experiments indicate that domain overfitting becomes more dominant with increasing amounts of adaptation data. While more adaptation data monotonically improves the model's accuracy for the target domain, it may degrade the performance on other domains. This may limit the applicability of the model for unseen real-world data. Our models were evaluated on different test sets from several domains to obtain a reliable estimate of the real-world performance and avoid selecting models that suffer from domain overfitting.

Overall, our experiments in the different studies show that substantial relative improvements in speech recognition for the target domain can be achieved with comparatively little training data. In the case of a domain mismatch, adaptation with a smaller subset may be more goal-directed than more training data. The main advantage of the fine-tuning approach is that an improvement in speech

recognition performance for the target domain can be achieved with comparatively little training time, i.e., without investing several weeks to months to train only one model. The fine-tuning approach allows training and evaluation of a multitude of models instead of only a few, which is advantageous both for research and adaptation for Audio Mining users' demands.

5.6.2 List of Contributions

List of scientific contributions in this chapter:

- Two-staged LF-MMI acoustic model domain adaptation was proposed and investigated, combining data augmentation for acoustic robustness with acoustic model fine-tuning. The approach was studied using a leave-one-speaker-out cross-validation.
- Automatic transcript alignment was studied for semi-automatic generation of in-domain, oral history adaptation data.
- Investigation of the influence of the amount of adaptation data on the general robustness and domain overfitting were performed.
- Domain overfitting through fine-tuning within different oral history subdomains was studied with adaptation data from two German oral history archives.

6 Multi-Staged Cross-Lingual Acoustic Model Adaptation

In the previous chapters, we first investigated and improved the acoustic robustness for the challenging recording conditions of oral history interviews. Then we investigated adaptation via fine-tuning on automatically transcript-aligned speech to adapt the acoustic model to the oral history target domain. Data augmentation and fine-tuning were reasonably combined to improve the recognition performance for the oral history domain substantially.

However, the adaptation via fine-tuning can lead to domain overfitting for some test sets, decreasing the models' overall robustness and applicability for real-world systems. In this chapter, we propose and study a multi-staged, cross-lingual adaptation to overcome this limitation and further improve the acoustic model robustness for different domains. The proposed approach aims at utilizing training data from heterogeneous domains both in the same and different languages. Particularly, we exploit publicly available English speech recognition corpora from diverse domains for training.

Automatic transcription of oral history interviews is not only of interest for German. As discussed in Section 3.3.1, it is also a relevant research topic for many low-resource languages for which generally substantially less annotated speech is available for training. This chapter also contributes to this low-resource speech recognition research, investigating cross-lingual adaptation from English to German with a small amount of German oral history training data for LF-MMI models.

The chapter is structured as follows. Section 6.2 provides an overview of related works on cross-lingual and multilingual adaptation in automatic speech recognition. Section 6.3 presents the proposed approach. The experimental setup for evaluation is presented in Section 6.4. Section 6.5 presents and discusses the results. In Section 6.6, we summarize the chapter's findings and contributions.

6.1 Thesis Author Contribution

Parts of this chapter are covered in the publication

> Michael Gref, Oliver Walter, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Multi-staged cross-lingual acoustic model adaption for robust speech recognition in real-world applications - A case study on German oral history interviews. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6354–6362. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.780

in which the multi-staged cross-lingual adaptation was proposed and studied by the thesis author. All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author. A co-author supported the experiments in coordination with the thesis author by training the English source model described in Section 6.4.1 using the noise and reverberation data augmentation, model, and training routines proposed and studied by the thesis author. Respective author contributions are given in Appendix C.

Section 6.3–6.5 summarize and extend the paper's content. The experiments, analysis, and discussions in the paper were extended and continued by the thesis author in the presented research work to put the overall contribution of the proposed approach into perspective to the other chapter's experiments, findings, and contributions.

6.2 Related Work

Cross-lingual acoustic model adaptation or *knowledge transfer* in automatic speech recognition aims at utilizing the knowledge of models trained in one language to improve recognition performance for a target language. The general idea is to exploit similarities between two languages by training a system on large amounts of annotated speech in the source language and adapting the system to the target language with usually only little annotated speech. A related task is *multilingual speech recognition*, where a system is trained simultaneously on multiple languages. Cross-lingual adaptation and multilingual training are often applied for *low-resource languages* where only insufficient training data is available to train single-language systems successfully. If the goal is to improve recognition performance for one particular target language, cross-lingual adaptation is often applied on top of multilingual acoustic models.

For DNN acoustic models, the approaches share the idea that the lower layers of the network are primary language-independent while the layers near the output are language-dependent. For multilingual trained systems, this is due to the simultaneous training on multiple languages, e.g., as studied by Huang et al. [2013], Ghoshal et al. [2013], and Grezl et al. [2014] for cross-lingual adaptation of multilingual DNN-HMM acoustic models with shared hidden layers. For cross-lingual

systems adapted from one language to another, this is the case for related languages with similar characteristics. For instance, this is studied by Chuangsuwanich et al. [2016], who investigated automatically identifying language subsets of multilingual training data close to the target language and more beneficial for adaptation.

In the late 2010s, cross-lingual adaptation still was a recent research topic enabling substantial improvements for low-resource speech recognition tasks. For instance, Xu et al. [2016] studied semi-supervised learning and cross-lingual knowledge transfer with multilingual data and neural network fine-tuning. Feng and Lee [2018] investigated cross-lingual knowledge transfer in a multilingual setup with language-dependent pre-final layers under each softmax output layer.

Only a few works studied cross-lingual adaptation for sequence-discriminative LF-MMI hybrid acoustic models. Ma et al. [2017] studied multilingual training using LF-MMI models with a joint LF-MMI output layer across languages followed by the adaptation to a low-resourced target language. Pulugundla et al. [2018] studied and proposed a multilingual LF-MMI trained system for low resource Indian language speech recognition.

Near the end of the 2010s, the research focus is increasingly moving towards end-to-end speech recognition. Cross-lingual adaptation is of particular research interest for these models, as they are often simpler to train than hybrid systems using phonetic representations and combining HMMs and DNNs, cf. Section 2.3. For instance, Kunze et al. [2017] studied the cross-lingual adaptation from English to German for the end-to-end *wav2letter* model [Collobert et al., 2016]. The authors combined ten different German corpora, mainly from the Bavarian Archive for Speech Signals (BAS) [Schiel, 1998], with overall 416 hours for adaptation. Cross-lingual adaptation of a CTC-based multilingual end-to-end acoustic model was studied by Tong et al. [2018]. The authors further investigated cross-lingual adaptation with the end-to-end implementation of the LF-MMI criterion (cf. Section 2.3.2) in [Tong et al., 2019]. In recent years, Vyas et al. [2021] investigated CTC and end-to-end LF-MMI training with the recent wav2vec 2.0 model [Baevski et al., 2020] for training an English system and for cross-lingual adaptation from English to Tagalog and Swahili. Luo et al. [2021] studied adaptation with an English-trained *QuartzNet* end-to-end system [Kriman et al., 2020] for cross-lingual adaptation to German, Spanish, and Russian, adaptation to different English accents, and other domains within the English language. The authors also studied cross-lingual adaptation from Mandarin to Cantonese.

Most of these related works perform only a single-staged (supervised) adaptation from multilingual or single-language source models to the target language. In our approach, we study a multi-staged adaptation for LF-MMI models not only to perform language adaptation but particularly to improve the model for the German oral history domain and simultaneously minimize domain overfitting.

6.3 Proposed Approach

The proposed multi-stage cross-lingual adaptation consists of three subsequent training stages, as shown in Figure 6.1. We first use a vast amount of data from a different language to pre-train the acoustic model and then train in two stages on decreasing data for more nearby domains in the target language. At the transition of each stage, we transfer the corresponding learned knowledge to the initial network in the next stage. The approach is designed for the specific use case to utilize three vastly differently sized data sets from different domains. For our use case, we consider several combined English corpora in the first stage, German broadcast data in the second, and only a small adaptation set of the German oral history target domain in the last adaptation stage.

6.3.1 Stage 1: Other-Language Pre-Training

In the first stage of the proposed approach, a robust acoustic model is pre-trained using a vast amount of different-language, heterogeneous, out-of-domain training data. Generally, a model reasonably trained on such data learns to perform a robust extraction of relevant acoustic input features and learns useful internal representations for the classification task. We assume that these aspects are, at least to some extent, language-independent for related languages. We then apply a weight transfer of all hidden layers and use these layers to initialize the acoustic model neural network training in the second stage to use this knowledge for tasks in languages with less available data.

English is probably the language with the most available training data for speech recognition. Therefore, we propose combining several English corpora from different domains to train the acoustic model in this stage. By combining corpora from different domains, we obtain a heterogeneous training set covering a wide range of different conditions.[1] Furthermore, we apply the three-fold noise and reverberation data augmentation from Section 4.4 in our approach to improve the model's robustness and generalization. In this and all subsequent stages, we again apply Kaldi's default speed perturbation [Ko et al., 2015] to further increase the variability and amount of data three-fold using constant speed factors 0.9 and 1.1. Thus, the training data in this stage is increased nine-fold.

The LF-MMI acoustic model training configuration is the same as for the German models in the previous chapters. As a default step of the acoustic model training, we train an i-vector extractor on English data in this stage.

[1] The combination of multiple public English speech corpora for ASR training was also proposed and studied by Chan et al. [2021] for the Google SpeechStew ASR system after our publication [Gref et al., 2020].

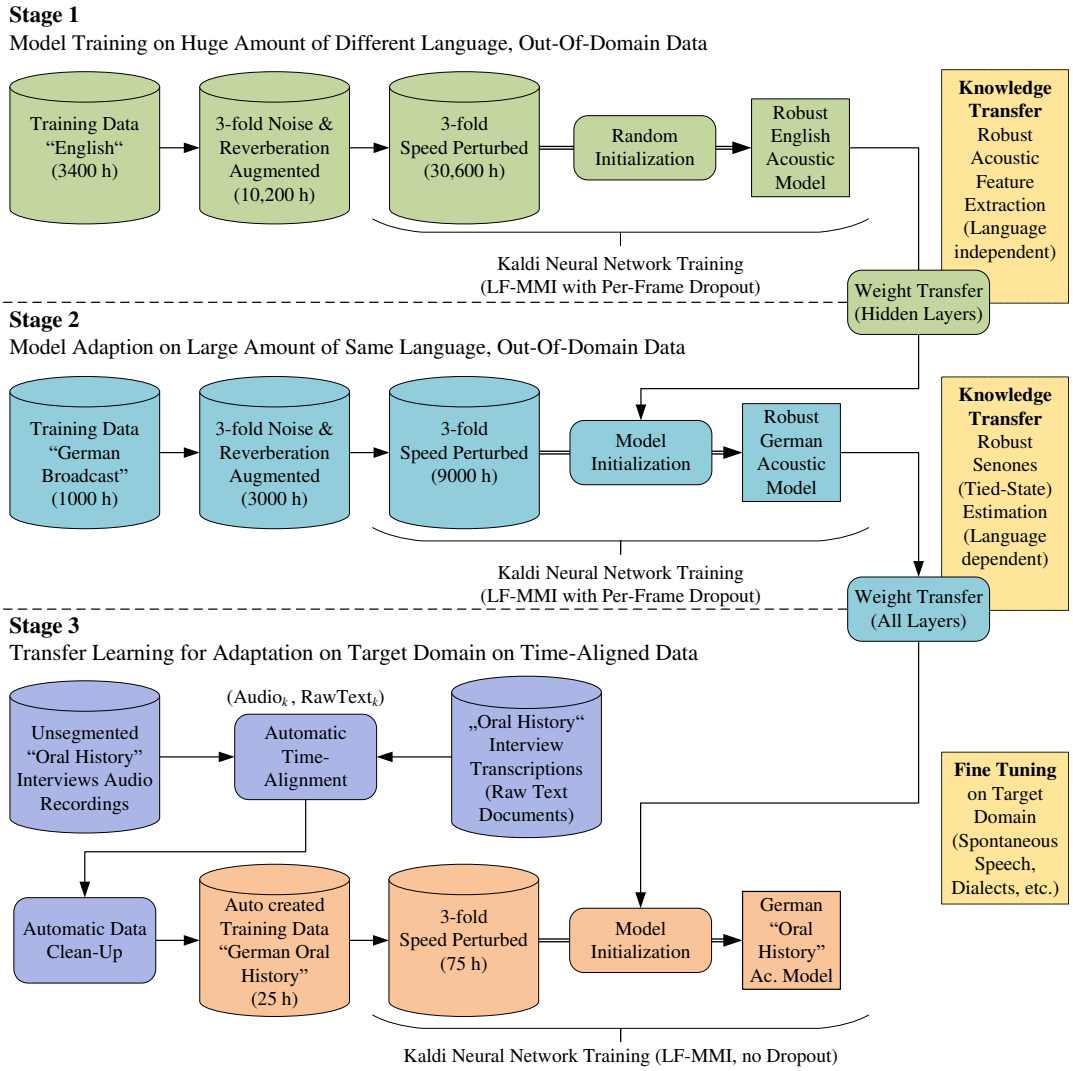


Figure 6.1: Proposed cross-lingual, multi-staged acoustic model adaptation approach. We first pre-train the acoustic model on 9×3400 hours of English speech data, then adapt it to the German language using 9×1000 hours. Finally, the acoustic model is adapted to the German oral history target domain with 3×25 hours of automatic transcript-aligned speech. Three-fold noise and reverberation data augmentation is applied for English and German Broadcast training to improve the model's acoustic robustness. In all stages, Kaldi's default speed perturbation is applied.

6.3.2 Stage 2: Same-Language Cross-Lingual Adaptation

In the second stage, the acoustic model is adapted to the language of the target domain, however, utilizing a large amount of training data from another domain. We perform the cross-lingual adaptation implicitly and apply no phoneme mapping. We replace the language-dependent LF-MMI and CE-regularization output layers of the LF-MMI acoustic model trained on English with randomly initialized output layers for tied states of German bootstrap GMM-HMM models—and then train this LF-MMI acoustic model on German speech.

In this stage, the feature extraction and representation learning of the lower layers is further improved for the target language while the classification of language-dependent subphonetic units is learned. As in the previous stage, training data is increased nine-fold to improve acoustic robustness and generalization as in the previous stage using noise and reverberation data augmentation and speed perturbation.

We apply the English-trained i-vector extractor from the previous stage for the German acoustic model training in this stage. We do not adapt the extractor on data from the target language as we consider the i-vectors to be mainly language-independent.

6.3.3 Stage 3: Same-Language Domain Adaptation

The acoustic model is adapted to the target domain in the last stage. We not only transfer the hidden layers in this stage but utilize a full weight transfer. This adaptation is the fine-tuning approach investigated in Chapter 5. In contrast to the previous stage, we do not replace the output layer since we use the same phone set and phonetic decision tree in Stage 2 and Stage 3. We obtain the training lexicons using the same grapheme-to-phoneme (G2P) pronunciation model.

Since we lack training data for German oral history interviews, we again utilize the adaptation data created with the automatic audio transcript alignment investigated Section 5.4 for the same-language domain adaptation.

6.4 Experimental Setup

In the following section, we describe the experimental setup for our investigation of the proposed multi-stage cross-lingual adaptation. The experimental setup is in large parts the same as in the previous chapters to ensure comparable results. In particular, we again carry out all experiments using the Kaldi ASR toolkit [Povey et al., 2011] with LF-MMI training [Povey et al., 2016] of the TDNN-LSTM acoustic model architecture selected in Section 4.3.

However, in this chapter, we use a slightly updated version of the training routine compared to the experiments in Chapter 4 and the first two of three adaptation studies in Chapter 5. In these experiments, we used the training routine of the original 3-fold (v1) acoustic model investigated and proposed in Section 4.4. The training routine we apply in this chapter was improved to train an improved *3-fold v1.1* baseline model and varies in some details, cf. Section 4.4.5. In particular, for the bootstrap GMM-HMM model training, fixed relative ratios of segments from the entire training data set are used instead of fixed amounts of segments, enabling the training routine to handle differently sized training data automatically. This is particularly useful for the experiments in this chapter, where the amount of training data of each stage varies substantially. Furthermore, the training is slightly improved by using more heterogeneous data for the i-vector extractor training and a slightly adjusted learning rate scheduling. In the following, we present and discuss the remaining experimental setup details for each stage and different types of experiments.

6.4.1 Training of English Model in Stage 1

For English model training in Stage 1, we combined the training data sets from the well-known corpora *Librispeech* [Panayotov et al., 2015], the 2018s, 240-hour version of the English *Common Voice Corpus*[2], *Switchboard* [Godfrey et al., 1992], and *Fisher* [Cieri et al., 2004]. Overall, the English training data comprises more than 3400 hours of annotated speech. We create two distorted versions in addition to the clean data set using the noise and reverberation data augmentation from Section 4.4. The first version uses a random 5–10 dB signal-to-noise ratio, the second one 10–20 dB. We utilize 266 room impulse responses of small and medium-sized rooms for reverberation and several noises recorded in real-life scenarios in both versions.

For a preliminary assessment of the acoustic model, an English general-propose language model for decoding based on crawled texts is used. The English model achieves a 9.17 % word error rate on Librispeech test-clean and 18.30 % on the data from *Voices Obscured in Complex Environmental Settings* (*VOiCES*) [Richey et al., 2018].

6.4.2 Adaptation to German in Stage 2

For the cross-lingual adaptation from English to German in Stage 2, we again utilize the 1000-hour large-scale corpus of German broadcast speech *GerTV1000h* also used in the previous chapters for acoustic model training. For the noise and

[2] https://commonvoice.mozilla.org

reverberation data augmentation in this stage, we apply the same 3-fold setup as in Section 4.4.3. In summary, two additional distorted versions of the clean training are created. The first version applies artificial reverberation using 266 room impulse responses of small and medium-sized rooms. The second one is created with both reverberation and noises using a random 10–20 dB signal-to-noise ratio.

Related works indicate that hyperparameter tuning, such as layer-freezing or a decreased learning rate for the lower layers, can improve cross-lingual adaptation. However, most works studied the cross-lingual adaptation with comparably small adaptation set sizes. Due to the week-long training of one LF-MMI acoustic model on 9000 hours of annotated speech, it is not feasible to perform hyperparameter tuning in the presented research work. Therefore, we use the default training routine as for *from-scratch training* with four epochs, default learning rates, and per-frame dropout, cf. Table 5.1. We show that the cross-lingual adaptation improves robustness and decreases domain overfitting with this initial configuration.

6.4.3 Adaptation to Oral History Domain in Stage 3

The adaptation of the German broadcast model from Stage 2 to the oral history domain is based on the extensively studied acoustic model adaptation from Chapter 5. We use the best configuration investigated in that chapter, especially the default learning rate, which has been proven beneficial in several different experiments (cf. Sections 5.4.3 and 5.5.2). Furthermore, we also do not apply per-frame dropout for this adaptation stage. We utilize the 25-hour $\mathrm{OH}_{150}^{10\%}$ oral history adaptation set of automatically transcript-aligned oral history interviews of 150 different speakers, cf. Section 5.5.

6.4.4 Evaluation

The evaluation is performed on the same three test sets from the broadcast domain and the three test sets from other domains as in the previous experiments, cf. Section 3.4. We study the influence of the different adaptations on domain overfitting by comparing the relative improvements on the diverse domains. We do not perform adaptation or evaluation on the HdG data sets studied in Section 5.5. These data sets were created in 2021, while the presented experiments in this chapter were conducted in 2019 and published in 2020.

For decoding, we utilize the same default broadcast language model also used in most experiments in the previous chapters. Additionally, we evaluate all models with the large broadcast language model *Large LM* with a 2 million words vocabulary trained on 1.6 billion running words, cf. Section 3.5. Comparing the

recognition performance of the acoustic models with two different language models allows us to better assess the influence of cross-lingual adaptation by further reducing the influence of the language model on the evaluation.

6.4.5 Performed Experiments

We use two baseline models to compare our proposed approach. These models are trained without neural network initialization from a prior model, and no further adaptation is performed. The first baseline is the 3-fold v1.1 acoustic model trained from scratch on the 1000 hours of German broadcast data with the data augmentation and training setup described for Stage 2. The second baseline is a model trained from scratch on the 25-hour $\mathrm{OH}_{150}^{10\%}$ adaptation set of automatically transcript-aligned oral history interviews.

Furthermore, to determine how each stage contributes to the improvements of the proposed approach, we carry out three ablation study experiments in which we omit one of the stages from training. The ablation study without English pre-training Stage 1 is the adaptation approach from German Broadcast to the oral history domain studied in Chapter 5 but with 3-fold v1.1 as the source model. The experiment with removed Stage 3 is conventional (single-staged) cross-lingual adaptation as referenced in the related work. In the ablation study without Stage 2, the English model is cross-lingually adapted to the oral history domain using the 25-hour adaptation set. This experiment contributes to low-resource challenges by investigating the effects of cross-lingual adaptation with a small data set from the target language for different domains. Thus, overall, we compare the proposed approach with overall five different models.

In experiments without English pre-training—both in the baselines and in the ablation study—the i-vector extractor is also trained with the corresponding data used in the first respective training step. We also study the influence of the i-vector extractor on the knowledge transfer from English pre-training Stage 1 by comparing the German and English trained i-vector extractor in an additional experiment.

6.5 Results and Discussion

The results of the proposed approach, baseline models, and the ablation study experiments are summarized in Table 6.1 for both language models. We have not included the detailed results on the four individual DiSCo subsets in the table for the sake of clarity. The results on the individual subsets show the same trends as DiSCo Average and do not allow to draw further conclusions. For the sake of completeness, the results are added in Table B.4 in the Appendix.

Table 6.1: Multi-staged cross-lingual adaptation results compared to two baselines and three ablation studies. Results are reported for the default and the large decoding language model as word error rates in percent. The upper row per test set shows the word error rate decoding with the default language model. The respective next row (*+Large LM*) shows the results of the acoustic models on this test set with the larger language model.

	Baseline		**Ablation Study**			
	Broadcast 3-fold v1.1	**Oral History**	**Removing Stage 1**	**Removing Stage 2**	**Removing Stage 3**	**Proposed Approach**
Stage 1 (English)				×	×	×
Stage 2 (German Broadcast)	×		×		×	×
Stage 3 (Ger. Oral History)		×	×	×		×
GerTV Dev Set	13.6	22.5	13.8	18.4	**13.4**	13.7
+Large LM	**12.9**	22.7	**12.9**	18.1	**12.9**	13.1
DiSCo Average	**11.9**	27.9	12.5	20.8	**11.9**	12.4
+Large LM	12.2	28.7	12.5	20.8	**12.1**	12.4
German Broadcast 2016	11.7	21.1	**11.3**	15.9	11.8	**11.3**
+Large LM	9.9	20.2	**9.4**	14.3	9.7	9.6
Challenging Broadcast	19.7	33.6	19.8	26.6	19.5	**19.4**
+Large LM	17.4	31.9	17.6	24.6	**17.3**	17.4
Oral History	27.7	37.4	**25.9**	28.7	27.4	**25.9**
+Large LM	27.1	38.2	25.3	28.5	26.5	**25.2**
Interaction	48.2	69.1	48.2	58.6	47.4	**47.1**
+Large LM	51.2	72.0	51.4	60.1	50.6	**50.3**
Spoken QALD-7	19.0	36.7	19.1	31.1	18.6	**18.4**
+Large LM	14.8	30.1	14.3	24.6	13.8	**13.6**

Overall, Table 6.1 shows that the multi-staged cross-lingual adaptation improves speech recognition and robustness, particularly for the three test sets from non-broadcast domains. The proposed approach achieves the best results for all these three test sets—consistently with both language models. For the broadcast domain, the results are less evident. In the following, we discuss the individual results in detail compared to the different reference models and, in particular, the influence of the three different stages.

6.5.1 Comparison to Baselines

The oral history baseline performs significantly worse than the broadcast baseline on all sets—even on the target domain. This is due to the considerably smaller amount of training data. Compared to the broadcast baseline, the proposed approach achieves a relative word error improvement of 6.3 % on the oral history test set using the default language model. The relative improvement using the larger language model is 7.1 %. For Interaction, the relative improvement is less substantial in the range 1.9–2.3 % depending on the language model. For Spoken QALD-7, the word error rate also improves with the proposed approach with both language models. However, the improvements are somewhat different for the two language models. A relative improvement of 2.9 % is achieved with the Default LM. With Large LM, the improvement is 8.2 % relative to 3-fold v1.1. As discussed in Section 4.4.5, the large language model's vocabulary is better suited than the default language model for modeling the various entities in Spoken QALD-7's test prompts. Therefore, we assume that the poor modeling properties of the default language model for this test set tend to mask the improvements of the acoustic model achieved by the multi-stage cross-lingual adaptation.

The proposed approach simultaneously improves or maintains the recognition performance on Challenging Broadcast and German Broadcast 2016 with both language models. However, we observe a decrease in performance for GerTV Dev and DiSCo Average. We consider the reduced performance is because this data is already very close to the conditions presented by the GerTV1000h broadcast training used to train the broadcast baseline model. The following section further explores how the three stages contribute to these observations.

6.5.2 Ablation Study

We investigate the contribution of each of the three stages to the recognition performance of the proposed approach by removing one of three stages at a time. The overall word error rates for each test set with both language models are summarized under *Ablation Study* in Table 6.1. In the first ablation study setup, we remove the English pre-training of the acoustic model in Stage 1. In particular, we

randomly initialize the acoustic model in Stage 2 for training on German broadcast data from scratch and then adapted it to the oral history domain in Stage 3. In the second setup, we omit the adaptation to German broadcast in Stage 2. Instead, we adapt the English model from Stage 1 to the target language and domain in one stage using only the 25-hour oral history adaptation set created with automatic transcript alignment. In the last setup, we omit the fine-tuning to the oral history target domain and evaluate the model adapted from English to German with 3×1000 hours of broadcast speech in Stage 2.

Undoubtedly, training on 3×1000 hours of German language data in Stage 2 has the most significant impact on recognition performance on all German test sets. The impact of English pre-training in Stage 1 and adaptation according to oral history in Stage 3 on the different domains is not that evident, making a detailed comparison of the two worthwhile. Figure 6.2 shows the effect of removing the stage for all test sets when decoding with both language models. A clear difference is evident between test sets of the broadcast domain and the three test sets of other domains. Both stages contribute to an improvement in recognition performance for the other-domain test sets. Unsurprisingly, for Oral History, the domain adaptation in Stage 3 has the greatest impact. However, for Interaction and Spoken QALD-7, the cross-lingual in Stage 1 has a greater impact.

For the broadcast domain, the results are more heterogeneous. As presented in Chapter 5, adaptation to the oral history domain tends to deteriorate the performance on DiSCo due to domain overfitting. However, the cross-lingual adaptation from English to German in Stage 1 improves the performance for all configurations, except German Broadcast 2016 and GerTV Dev with large language models. In the following sections, we examine the contribution and impact of each stage in further detail by investigating the relative improvement of the Stages to the respective baseline.

Removing Stage 1: Removing English Pre-Training

To make the influence of the cross-lingual adaptation better comparable for the different test sets and language models, we visualized the improvement relative to the broadcast 3-fold v1.1 baseline of the proposed approach and the mono-lingual model with removed Stage 1 in Figure 6.3. The English pre-training Stage 1 has little influence on the Oral History test set, with an almost negligible improvement. However, we observe an improved performance for virtually all other test sets due to the cross-lingual Stage 1. For Challenging Broadcast and Interaction, the cross-lingual adaptation enables the system to improve instead of degrading results compared to the baseline. For DiSCo Average, the results are just minimally better, although the recognition performance with the baseline for this test set is still better. For Spoken QALD-7, we observe a quite substantial improvement

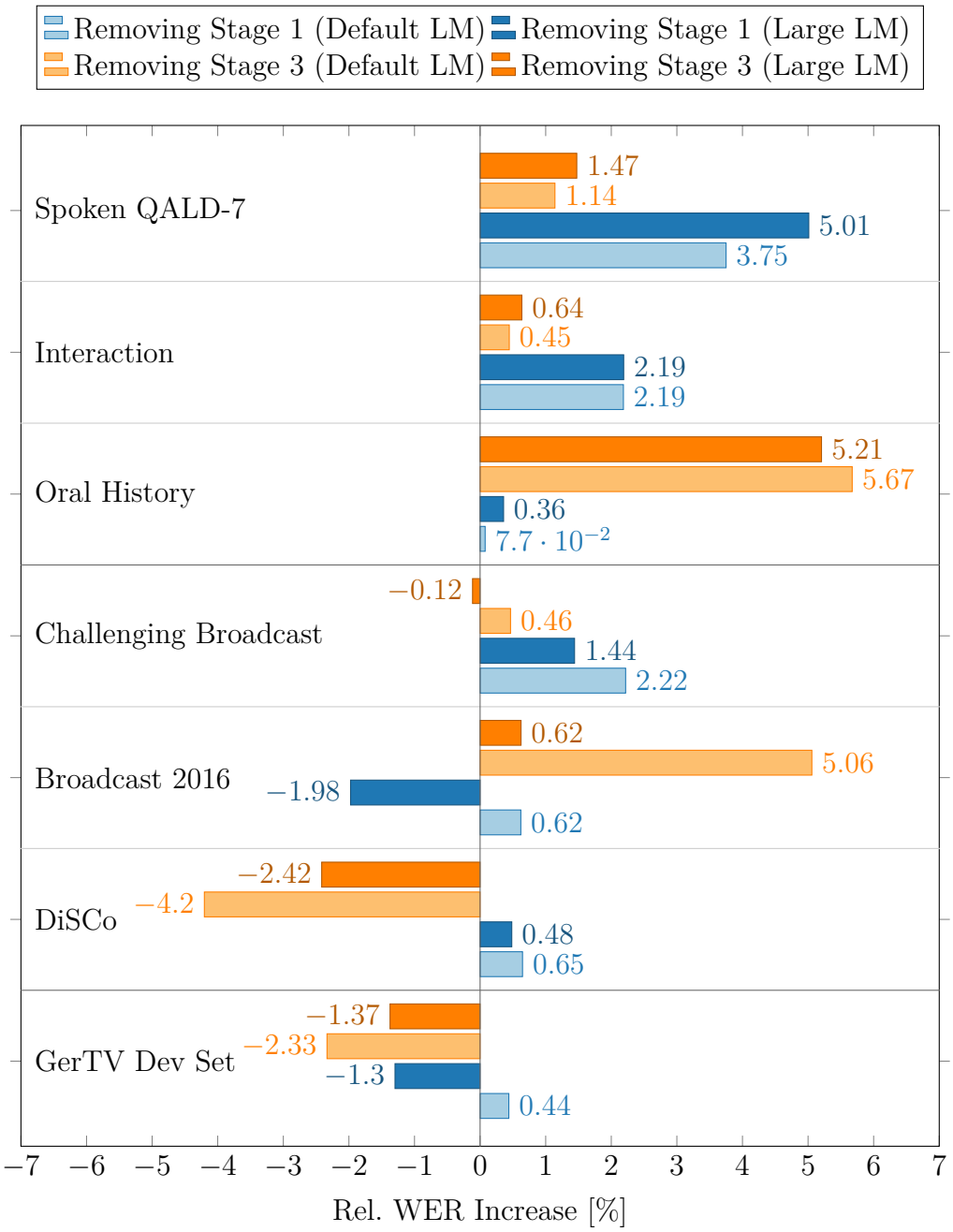


Figure 6.2: Ablation study of the multi-staged cross-lingual adaptation. The results are reported as an increase in the relative word error rate per test set and for two independent language models (LM) when a stage is removed from the proposed approach. An increase in the error rate indicates that the stage contributes to improving the recognition accuracy for the respective test set. Correspondingly, a decrease in the error rate indicates that the stage is detrimental to the approach for that test set and degrades recognition accuracy.

due to cross-lingual Stage 1. The previously discussed language model dependence and greater relative improvement with Large LM on this test set are revealed by the relatively large deviation from the diagonal axis. Overall, we infer from the results that the cross-lingual component of our proposed approach contributes to a consistent improvement on different domains and thus to a reduction of the domain overfitting that we observed in Chapter 5 in different mono-lingual adaptations.

Furthermore, the improvement achieved by the cross-lingual adaptation in the proposed approach is consistent for both language models on most test sets, except for German Broadcast 2016. The proposed approach still improves the recognition performance over the broadcast baseline for this test set. However, with the Large LM, the performance is better without cross-lingual pre-training, applying just a single-stage mono-lingual adaptation.

Adaptation from English to German Oral History with Removed Stage 2

As shown in Table 6.1, omitting training on the 1000 hours of German language data significantly and consistently degrades recognition performance for all test sets and domains. The results are also substantially worse than the broadcast baseline. However, comparing the results with the oral history baseline reveals that this single-stage cross-lingual adaptation is suitable for low-resource tasks where only a few hours of annotated speech for the language is available. Using only English training data and the small oral history adaptation set, we reduce word error from 37.4 to 28.7 % on Oral History decoding with the default language model. Results are similar for the large language model. This value is only one absolute percentage point higher than the result achieved by the robust 3-fold v1.1 broadcast baseline trained on 3×1000 hours of manually annotated German broadcast speech. This *low-resource-trained* model even outperforms the *Mix-models* from the data augmentation Chapter 4 on the Oral History test set using only one-fortieth training data (cf. Table 4.10). Thus, adapting from a rich-resourced language directly to the target language and domain is suitable if no other data is available for training in the target language. In particular, we show this for LF-MMI acoustic models with an LSTM-TDNN topology, which usually requires a lot of training data.

For real-world applications, one cannot expect this low-resource cross-lingual model to be as robust as models trained robustly on large-scale data from the target language—but substantially more robust than training from scratch on the small data set only. We observe a word error rate in the range 14.3–26.6 % for the broadcast domain. The range is 24.6–60.1 % for the other three domains. Figure 6.4 shows the improvement of this mono-stage mono-lingual adaptation relative to the from-scratch trained oral history baseline. By initializing with the

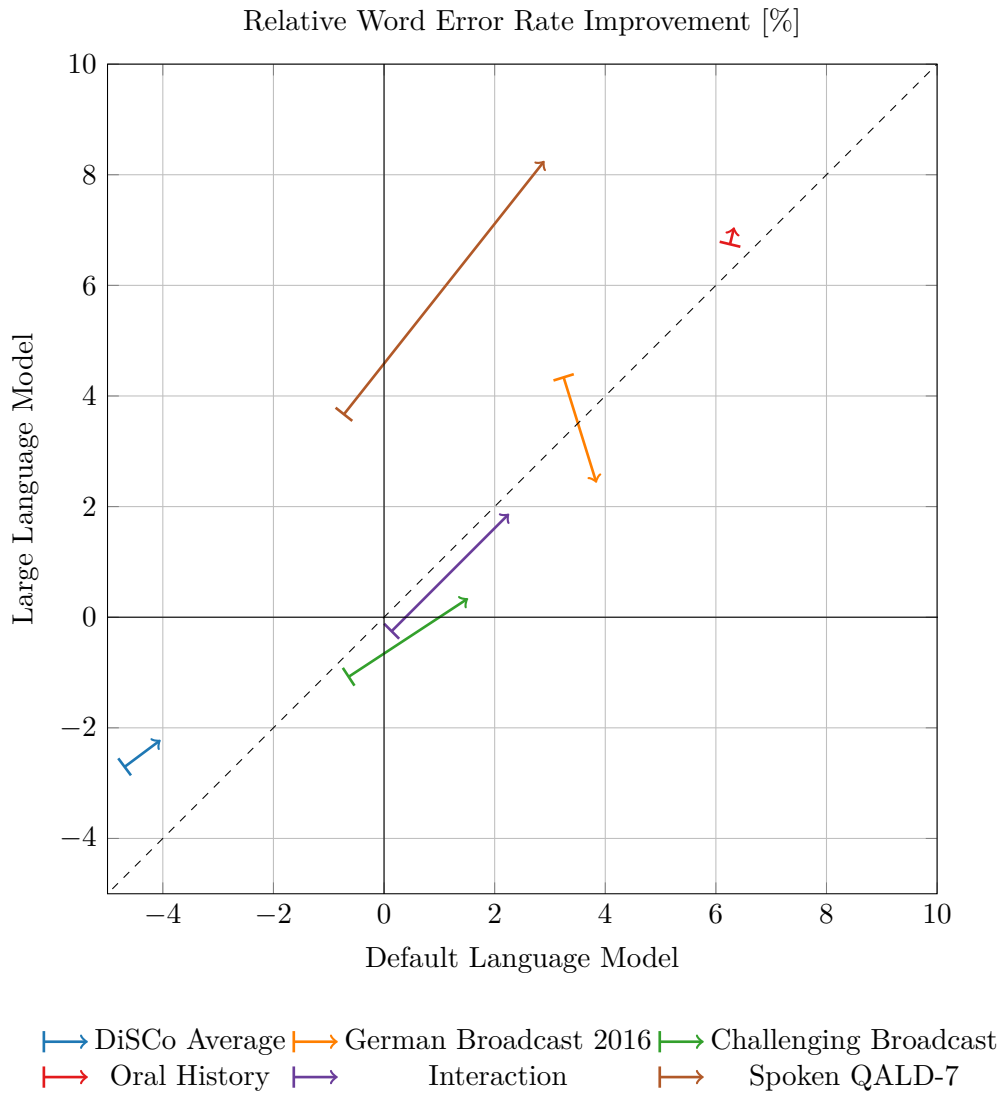


Figure 6.3: Relative WER improvement of the proposed multi-staged cross-lingual acoustic model adaptation compared to a mono-lingual adaptation, excluding the English pre-training Stage 1 for both language models. Each arrow represents a WER improvement for a test set relative to the broadcast 3-fold v1.1 baseline acoustic model. The starting point of each arrow represents the relative improvement of the mono-lingual adaptation (without Stage 1) from German Broadcast to Oral History. The end of the arrow represents the improvement of the proposed approach with multi-staged cross-lingual adaptation (including Stage 1). If the arrow points to the top right corner, the proposed approach simultaneously improves the word error rate for both language models. Values above zero indicate an improvement of the WER.

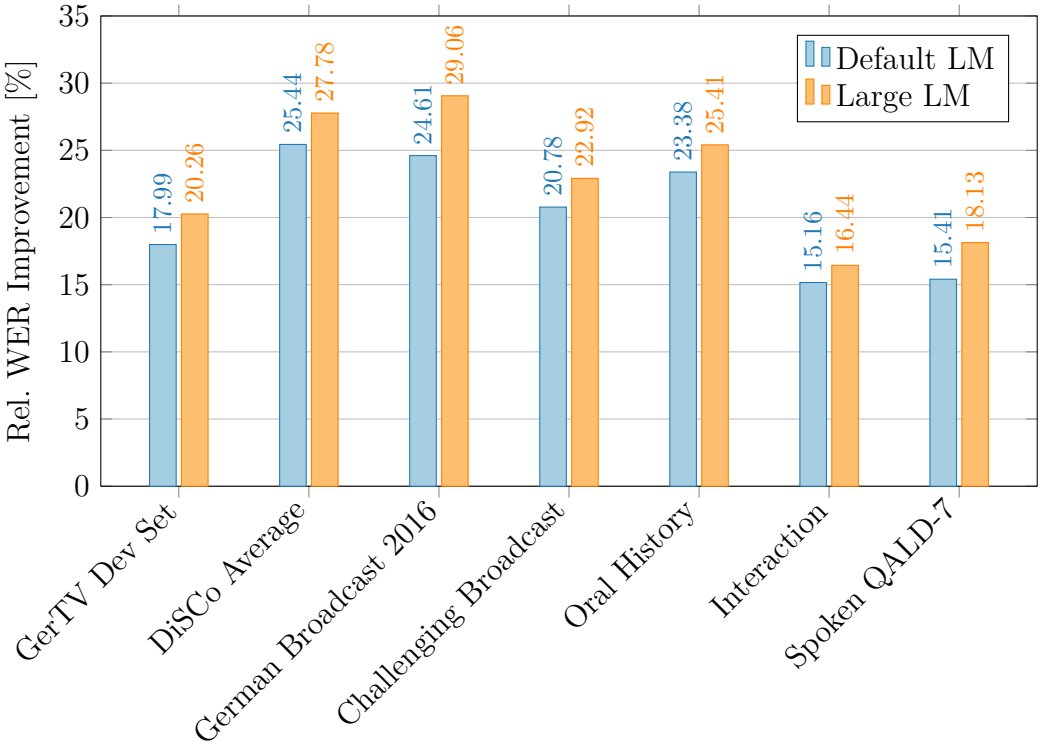


Figure 6.4: Improvements of single-stage cross-lingual adaptation from English to German with 25 hours of Oral History interviews relative to a from-scratch trained baseline.

English model, the word error rate for all test sets improves substantially between 15.2–29.1% relative. The improvements are consistent for all domains and best for broadcast and oral history. The results are very similar for both language models, with a slight advantage for the Large LM.

Adaptation from English to German Broadcast with Removed Stage 3

This experiment evaluates the initialization with the English-trained model for large-scale German speech corpora training compared to the broadcast baseline. In Figure 6.5, we show the improvement of the English pre-training relative to the 3-fold v1.1 baseline. The results are nowhere near as substantial as the previous experiment with the oral history adaptation and baseline. However, for the non-broadcast domains, we observe a small but consistent relative improvement on all three test sets with both language models. By far, the greatest improvement is achieved on Spoken QALD-7 with the large language model. The other improvements relative to the baseline are below or slightly above 2 %.

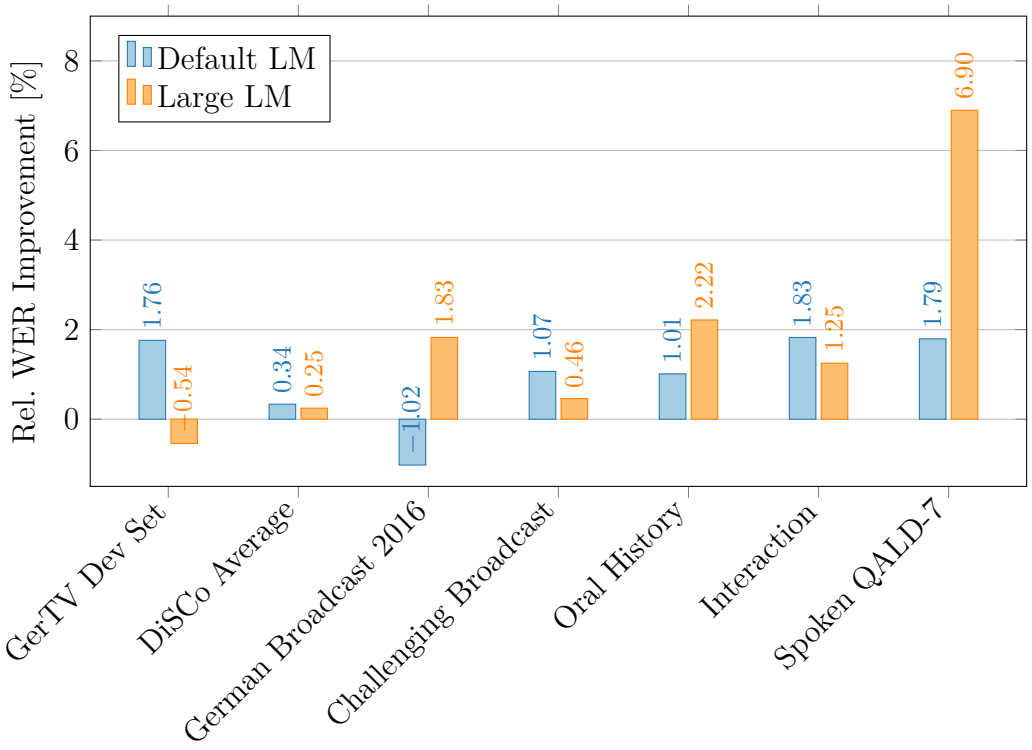


Figure 6.5: Improvements of single-stage cross-lingual adaptation from English to German with 3 × 1000 hours of broadcast speech relative to a from-scratch trained baseline.

6.5.3 Influence of the i-Vector Extractor Language

In the following experiments, we study whether the improved robustness through knowledge transfer of the English model is due to weight transfer, as we expect, or due to a better i-vector extractor trained on English data instead of German. Therefore, we compare each combination of German and English trained i-vector extractors with a random model initialization and initialization of hidden layers with the English trained model when training on the 1000 hour German broadcast data. For simplicity, we only report results on the small language model. The results are summarized in Table 6.2.

We achieved the best results on all test sets, except German Broadcast 2016, using the proposed setup with English-trained i-vector extractor and acoustic model initialization from the English-trained model. Using the German i-vector extractor with English model initialization leads to the worst result for Challenging Broadcast, Oral History, and Interaction. This is to be expected since i-vectors of the same speakers from two differently trained i-vector extractors point to different directions in the two different 100-dimensional vector spaces. Since the English

Table 6.2: Comparison of German and English trained i-vector extractors in acoustic model training on German broadcast data with random model initialization and with cross-lingual adaptation from English-trained model.

Acoustic Model Initialization	Rand.	Eng.	Rand.	Eng.
i-vector Language	Ger.	Ger.	Eng.	Eng.
GerTV Dev Set	13.6	13.5	13.5	**13.4**
DiSCo Average	**11.9**	**11.9**	**11.9**	**11.9**
German Broadcast 2016	11.7	**11.6**	11.8	11.8
Challenging Broadcast	19.7	19.8	19.6	**19.5**
Oral History	27.7	27.8	27.7	**27.4**
Interaction	48.2	48.5	47.8	**47.4**
Spoken QALD-7	19.0	19.2	19.4	**18.6**

model is trained with i-vectors from one space, using different i-vectors in the second stage causes wrong estimations of speakers, and this relation has to be relearned for the new vector space in Stage 2.

Using the English i-vector extractor instead of the German one with random acoustic model initialization leads to similar results. Only for Spoken QALD-7, the German i-vector seems to perform better with random acoustic model initialization. And for Interaction, this is the case for the English-trained i-vectors. Therefore, we infer that the proposed cross-lingual adaptation leads to improved acoustic models and not a better i-vector extractor trained on English data. In order to perform adaptation sensibly, the i-vector extractor must be used, which was also used to train the original model. The influence of the language, language combination, and amount of training data for the i-vector extractor training for LF-MMI acoustic models is further explored in the subsequent work by Wang et al. [2021].

6.6 Summary and Contributions

6.6.1 Summary

In this chapter, we proposed and investigated a multi-staged cross-lingual acoustic model adaptation approach to improve the acoustic model's robustness and decrease domain overfitting. Our approach addresses challenges where only lit-

tle training data for the target domain is present. It enables the exploitation of large-scale training data from other domains in both the same and other languages.

We studied our approach for our German oral history use case with intending to obtain an acoustic model that is robust enough to be applied in real-world applications. We first trained a robust acoustic model for English with more than 3000 hours of data. Then we adapted it to German using 1000 hours of German broadcast data. Three-fold noise and reverberation data augmentation from Chapter 4 was utilized in both stages. This model is again adapted using 25 hours of German, in-domain oral history interviews.

We performed extensive experiments to determine the robustness and real-world performance of the model. We evaluated the model not only with in-domain oral history data but also with our several German test sets from other domains and two different decoding language models. To thoroughly determine which stage of the proposed approach contributed to the improvements in the different domains, we conducted ablation study experiments. Thereby we have shown that the direct adaptation of LF-MMI acoustic models from one language to another leads to good results, even using only very little training data from the target domain. Thus, this observation can contribute to the ongoing research on speech recognition for under-resourced language.

The model trained with our proposed approach achieves a relative reduction of the word error rate by more than 30 % compared to a model trained from scratch only on the target domain, and 6–7 % relative compared to a model trained robustly on 1000 hours German broadcast training data. Overall, we achieve a 25.9 % word error rate on the Oral History test set with our default language model and 25.2 % with a larger language model, simultaneously improving the performance in the Interaction, speech assistant, and challenging broadcast domain.

6.6.2 List of Contributions

List of scientific contributions in this chapter:

- Multi-staged cross-lingual adaptation was proposed and investigated that reduces domain overfitting and increases the robustness of the domain-adapted LF-MMI acoustic model with a cross-lingual pre-training stage.

- Contributions to low-resource oral history speech recognition tasks were made by studying cross-lingual adaptation from English to German Oral History with only 25 hours of annotated speech achieving performance on the target domain similar to a from-scratch trained model on 1000 hours German broadcast speech.

- Investigation of the influence of the training data language was performed to study the influence on the i-vector extractor for LF-MMI acoustic models in cross-lingual adaptations.

7 Conclusion and Outlook

7.1 Conclusion

Automatic transcription of oral history interviews offers a variety of benefits to researchers and historians. It can significantly reduce workload and enable new types of analysis. However, the last 20 years of research have revealed various challenges that make the transcription of oral history interviews demanding. Even in recent years, nearly all works are characterized by a high word error rate—particularly when compared to other standard speech recognition benchmarks.

The present research dealt with developing a robust speech recognition system for German oral history interviews from 1980–2010 of the Deutsches Gedächtnis archive at the University of Hagen. A lack of representative training data for the oral history speech domain and a high word error rate of 55 % of the baseline system characterized the beginning of the presented research work in 2017. The main objective of this research was to develop and improve automatic speech recognition systems for the German oral history interviews of this archive. To achieve this goal despite the lack of data, we explored the adaptation of data and models described in the following. In our investigations, we particularly focused on the real-world performance of the system and investigated the possible overfitting to the target domain for all proposed systems.

First, we analyzed the challenges of oral history interviews in general, as reported in the literature, and examined the challenges of the German interviews we studied. For this purpose, we conducted several preliminary studies to investigate the respective challenges of the interviews and identify the components of the speech recognition system that need to be improved the most. Furthermore, we studied human transcription accuracy for interviews and postulated a human word error rate for transcription of interviews with high recording quality. We examined the speech rate of oral history interviews compared to speech recordings from other domains. Additionally, we investigated the language model and vocabulary influence using intrinsic evaluation metrics. We identified the acoustic model as the component to be improved, particularly the acoustic recording conditions of the interviews, characterized by room reverberation and noise, and the spontaneity in the interviewees' speech.

After comparing selected acoustic models, we investigated and compared methods for improving acoustic robustness. In our interviews, we identified room rever-

beration as one of the main acoustic challenges generated by a relatively large distance between the speakers and the microphone in small and medium-sized rooms. The robustness of the acoustic model was improved significantly using noise and reverberation data augmentation, without the need to increase the training data size. The performance not only improved for the oral history interviews but also for broadcast, speech assistance systems, and conversational interaction. With a three-fold increase training data set, an even more substantial improvement was achieved—however, at the expense of the training time of the model. This 3-fold model was trained on 3×1000 hours of broadcast speech and required over two months of training on our GPU cluster. It achieved a 28.2 % word error rate on oral history and concurrent new benchmarks on the test data of the other domains. Due to its robustness, it has been used in the Fraunhofer IAIS Audio Mining system as a default model, e.g., at the ARD. Furthermore, we have investigated different speech enhancement approaches as pre-processing for the ASR. Three different approaches for noise suppression worsened the error rate instead of improving it. However, the common method WPE served as a useful complement and improved the error rate to 27.9 %.

The enormous training time of the model and the lack of representative oral history data for training remained a challenge. We investigated acoustic model adaptation via fine-tuning to address both issues and adapt models more quickly to the target domain. In three studies, we explored different properties and facets of adaptation. In the first of the three studies, we investigated the combining data augmentation of the source model training and subsequent domain adaptation. A leave-one-speaker-out cross-validation experiment improved the average word error rate to 26.6 % using only 3.4 hours of adaptation data. The adaptation also improved the robustness for test data from other domains.

In the second study on acoustic model domain adaptation, we investigated automatic transcript alignment as an approach to semi-automatically generate adaptation data from transcribed but temporally unaligned oral history interviews using the robust 3-fold model. We showed that this type of semi-automatic data generation is purposefully for adaptation. We investigated the impact of the adaptation data size and learning rate on domain adaptation and domain overfitting. The best adaptation on 250 hours improved recognition performance on oral history to a 24.5 % word error rate. However, this came at the expense of the recognition performance deteriorating for specific other domains.

The third study on domain adaptation investigated domain overfitting in detail for oral history using more recently recorded oral history interviews with high recording conditions. We showed that the size of the adaptation data set has a significant impact on domain adaptation. If training and testing conditions do not overlap exactly, a smaller dataset is more appropriate to improve the model's

general recognition performance. Using an updated version of the 3-fold model (3-fold v2) as the source model and adapting it to 250 hours of speech, we achieved a 23.9 % word error rate for the interviews with mixed acoustic conditions mainly studied in the presented research work. The model's error rate is 16.1 % for the more recent oral history interviews with high recording quality. With adaptation on 31 hours of oral history interviews from two different archives, we achieved a good compromise of 24.6 % WER for oral history in mixed acoustic conditions and 15.7 % WER in clean acoustic conditions.

Further reduced domain overfitting is desirable for robust real-world performance for unseen data. For this purpose, we proposed a multi-stage cross-lingual domain adaptation of the acoustic model, exploiting the vast availability of different English corpora for the German speech recognition system to improve robustness. Using the proposed adaptation, we improved the robustness of the previous domain adaptation quite consistently for almost all different domains. In particular, the cross-lingual adaptation of the multi-staged approach improved the recognition performance for the domains that are not modeled by either the broadcast training data or the oral history adaptation data. This indicates a reduction in domain overfitting.

Overall, we trained the acoustic model to cope with the different, heterogeneous, and unpredictable challenges of oral history interviews through the various investigated and combined approaches in the presented research work. This significantly improved the robustness not only for this domain but simultaneously for other domains not seen during training. Figure 7.1 summarizes the progression of word error rates on the representative oral history test set in our work for the primary models. The overall best system achieved an error rate of 23.4 %. Overall, we have more than halved the error rate of the baseline system. By continuously evaluating the models and approaches on numerous datasets of different domains, we have ensured that the models work robustly for multiple unforeseen conditions and provide improved transcription accuracy. We achieved a transcription accuracy that allows automated indexing of large oral history archives via the Fraunhofer IAIS Audio Mining system and facilitates transcription for subtitling and further analysis by only requiring correction of transcription errors.

7.2 Outlook and Future Work

The tremendous research interest of the international community in automatic speech recognition yielded many new approaches in recent years. These approaches can be promising for future work on the robust speech recognition of oral history interviews to extend and further improve the results and methods of the present research work.

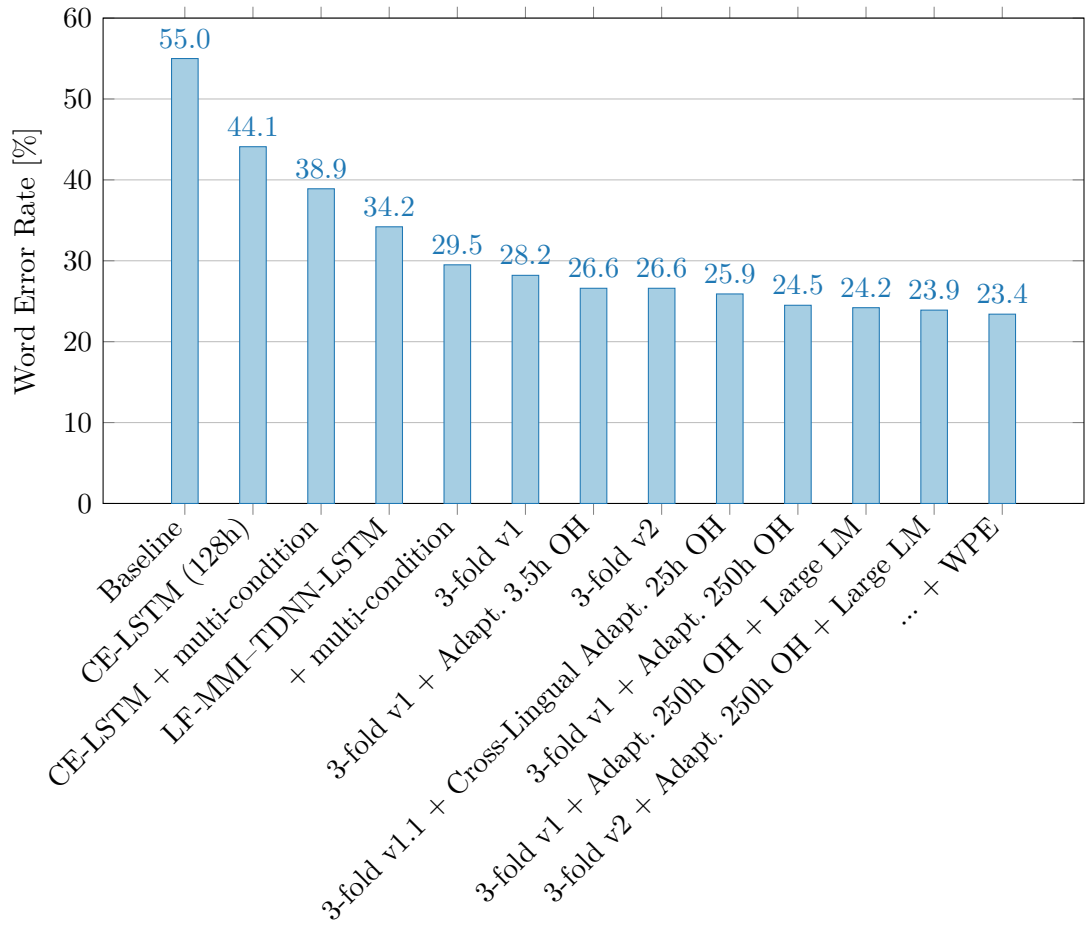


Figure 7.1: Summary of speech recognition results on the German oral history test set of the primary models and approaches studied in the presented research work.

The emerging research field of end-to-end speech recognition enables promising approaches to train models with substantially reduced explicit human knowledge. In particular, for spontaneous speech, dialects, and age- and health-related changes in the speech of contemporary witnesses, end-to-end approaches could be used for automated learning and improved modeling of pronunciations compared to explicit phonetic pronunciations from a lexicon. However, this is likely to require sufficiently large training data sets. The automated transcript alignment for adaptation investigated for oral history interviews in this thesis could be exploited to generate training data semi-automatically for oral history end-to-end systems.

Recent approaches, such as the wav2vec 2.0 [Baevski et al., 2020], utilize enormous amounts of non-transcribed speech for self-supervised pre-training of the system. These approaches can be promising for oral history interviews, as many

archives contain vast amounts of untranscribed interviews. For these systems, the noise and reverberation data augmentation investigated in this thesis, a multi-stage and possibly multi-lingual adaption can also be helpful to further improve the challenging task of automatic transcription of these interviews. The combination and comparison of these approaches with recent data augmentation approaches, such as SpecAugment [Park et al., 2019], and new, large-scale German corpora, such as CommonVoice, holds the potential for further substantial improvements.

Oral history interview archives are often very heterogeneous, with many unpredictable challenges regarding recording conditions and the recorded speakers. Robustly assessing the applicability of these systems for real-world applications and avoiding domain overfitting is crucial to developing systems that can be reasonably applied to transcribe these archives automatically. We advocate following the method proposed in our research work, evaluating and assessing the speech recognition systems not only on one oral history test set but on many different, well-curated, and documented data sets from diverse domains.

In this research work, substantial insights have been gained on the various challenges of oral history interviews for hybrid automatic speech recognition and the individual components. Methods for improvement have been proposed and investigated. These insights and investigated methods can be an important building block for future work on end-to-end speech recognition for oral history interviews to make the recognition performance more robust for this challenging domain.

A Appendix: Supplementary Toolkit and Software Descriptions

This appendix gives a supplementary overview of software and toolkits for automatic speech recognition and automatic transcription. Section A.1 presents and compares several relevant automatic speech recognition toolkits. Additionally, a detailed description of the toolkit *Kaldi*, which is used in this presented research work, is given. In Section A.2, the *Fraunhofer IAIS Audio Mining* system is presented, in which the speech recognition models trained in the presented work are integrated for real-world application.

A.1 Automatic Speech Recognition Toolkits

A.1.1 Overview

Like many other research branches of machine learning, automatic speech recognition research is primarily driven by flexible toolkits and frameworks that provide pre-implemented methods and algorithms for the efficient training of models. These toolboxes allow researchers to use state-of-the-art systems to study and extend new approaches and concepts.

Many toolkits for speech recognition have been proposed over the years and have been applied by many different researchers. One of the earliest toolkits used for speech recognition is probably the *Hidden Markov Model Toolkit* (*HTK*). As the name implies, hidden Markov models are the main focus of the toolkit. HTKs primary use certainly is speech recognition, even though the toolkit is not limited to this application—and it was state of the art for many years. The toolkit was first released to the public in the early 1990s. In 2015, the latest version of HTK was released in which Young et al. [2015] introduced deep neural networks to HTK.

Without claim to completeness to this list, other toolkits in the field of speech recognition that should be acknowledged (in chronological order) are *Julius* by Lee et al. [2001], *Sphinx-4* by Walker et al. [2004], and the *RWTH Aachen University Open Source Speech Recognition Toolkit* (RASR) by Rybach et al. [2009]. In 2011, Povey et al. [2011] proposed the *Kaldi speech recognition toolkit*. Kaldi has gained enormous popularity since its release. Deep neural networks were adopted in Kaldi very early, e.g., by Veselý et al. [2013]. Another noteworthy toolkit for

Table A.1: Number of citations of popular speech recognition toolkits at the end of October 2020. Numbers according to Google Scholar statistics.

Toolkit	Reference	Citations
HTK	HTK Book (in all versions since 1993)	6959
Julius	Lee et al. [2001]	783
Sphinx-4	Walker et al. [2004]	569
RASR	Rybach et al. [2009]	126
Kaldi	Povey et al. [2011]	4482
EESEN	Miao et al. [2015a]	579

speech recognition is *EESEN* by Miao et al. [2015a], a toolkit for end-to-end speech recognition using CTC.

Each of the aforementioned toolboxes deserves acknowledgment. For the present work, however, the question arises as to which of the toolboxes is the most promising and advanced to work with. It can be assumed that leading researchers use the most promising toolboxes for their work. Therefore, the number of citations in scientific publications of the toolboxes in question might give a good indication to answer the above question. An overview is given in Table A.1.

While all toolkits have a remarkable number of citations, Julius, Sphinx-4, RASR, and EESEN have not been applied in research to the same extent as HTK or Kaldi. With almost 4500 citations, respectively, nearly 7000 citations, Kaldi and HTK are far more popular in the research community than the other toolkits.

Considering that HTK was proposed two decades before Kaldi, the question remains open as to which toolbox is currently state of the art. In an attempt to answer this question, Figure A.1 compares the number of citations of both toolkits in recent years.

The number of citations for Kaldi has increased almost monotonously over the years since its publication in 2011. However, HTK's citations have been steadily decreasing since 2015. In 2016, Kaldi surpassed HTK in annual citations. It is not possible to give an undoubted reason for the enormous popularity of Kaldi over HTK. However, one of the main reasons certainly is the very fast adoption and constant improvement of deep neural networks for acoustic modeling in Kaldi since 2013—while they were adopted in HTK much later in 2015.

At the beginning of the present work in early 2016, the rapidly growing popularity of Kaldi was already foreseeable. Furthermore, at the Fraunhofer IAIS, Kaldi was already being applied in 2017—while EESEN was also studied, cf. Schmidt et al. [2016]. For these two reasons, and to work with the latest approaches in

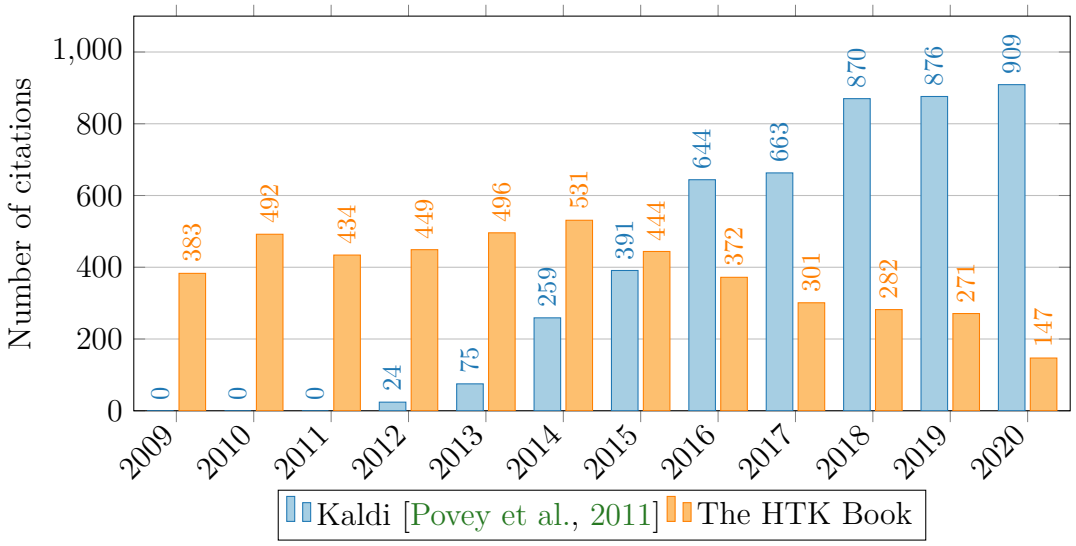


Figure A.1: Citation count of the Kaldi paper and the HTK book in recent years according to Google Scholar statistics in March 2022.

speech recognition throughout the present work, Kaldi was chosen as the toolkit for this work.

A.1.2 The Kaldi Speech Recognition Toolkit

According to Povey et al. [2011], the main focus of Kaldi is acoustic modeling research and the main important features of Kaldi that differentiate Kaldi from other toolkits, such as HTK, are:

- integration with (weighted) finite-state transducers using OpenFST [Allauzen et al., 2007]
- extensive linear algebra support
- extensible design
- open license
- complete recipes for widely known and available data sets
- thorough testing

The extensible design and open license of Kaldi enable many researchers to participate in developing new approaches that are often integrated quickly in the toolkit. New training routines and approaches are regularly integrated and provided as recipes for well-known data sets. This is of great value for the present

Table A.2: Components of conventional speech recognition systems as weighted finite-state transducers.[1]

FST	ASR Component	Input Symbols	Output Symbols
H	HMM	HMM States	CD Phones
C	Context Dependency	CD Phones	Phones
L	Pronunciation Lexicon	Phones	Words
G	Language Model (Grammar)	Words	Words

work, as these recipes can be used as a starting point for developing new speech recognition systems for German oral history interviews.

A.1.3 Speech Recognition with Weighted Finite-State Transducer in Kaldi

Weighted finite-state transducers (weighted FSTs or WFSTs) have become popular in speech recognition since they provide a natural representation of speech recognition systems' many components. WFSTs can represent hidden Markov models, context-dependency, pronunciation lexicons, language models, and alternative recognition outputs (*lattices*) [Mohri et al., 2002]. As finite-state transducers, each of these components has state transitions that are labeled with a weight, or *cost*, input and output symbols, as presented in Table A.2.

In order to illustrate how a pronunciation lexicon and a language model are represented as weighted finite-state transducers, simplified examples are shown in Figures A.2 and A.3. The lexicon has phoneme sequences as input and words as output. For different pronunciations of the same word, different pronunciation probabilities can be applied as weights. The language model input and output symbols are equal. Word sequence probabilities are modeled as weights for each state transition. While only a few nodes are shown in the examples, real finite-states transducers for large-scale vocabulary speech recognition systems comprise nodes and transitions for millions of recognizable words and word sequences. These transducers can barely be visualized.

The entire speech recognition decoding pipeline can be (simplified) represented by concatenating the transducer of each component in the following manner:

$$HCLG := H \circ C \circ L \circ G.$$

[1] Cf. http://www.inf.ed.ac.uk/teaching/courses/asr/2019-20/asr10-wfst.pdf, p. 8.

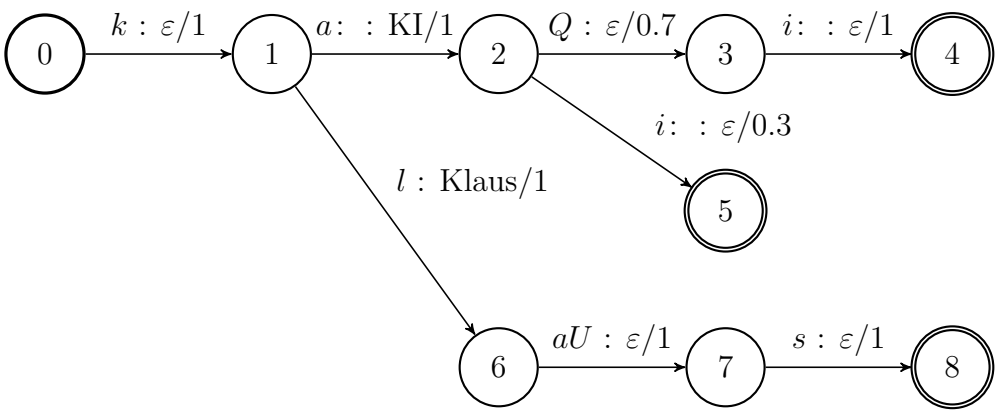


Figure A.2: Exemplary phonetic pronunciation lexicon with three entries as weighted finite-state transducer L. The WFST models the pronunciation of the German words *Klaus* and *KI*. The latter word is modeled with two alternative pronunciations: with and without a glottal stop between *a:* and *i:*. The labels $x : y/w$ at state transitions (arcs) mean x is an input symbol, y is an output symbol, and w is the respective weight for this state transition. The label ε means no input or output symbol for this transition. The entire input and output sequences are the concatenation of all respective symbols for a path from start node 0 to a final state (4, 5, and 8).

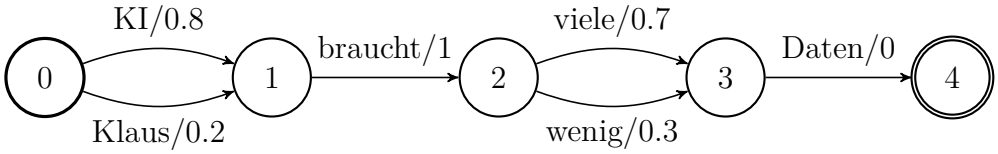


Figure A.3: Exemplary language model that models four possible word sequences as weighted finite-state transducer G. Since input and output symbols are equal at each arc, the words at each arc are only labeled once both for input and output.

This provides a conveniently and efficiently integrated weighted finite-state transducer with HMM states as input and words as output symbols [Mohri et al., 2002].

For numerical stability, negated logarithmic probabilities of the components, such as word sequence probabilities, pronunciation probabilities, and silence probabilities, are used as weights for the transducers, cf. Mohri et al. [2002]. Operations on the transducers weights are based on the *tropical semiring*

$$(\mathbb{R} \cup \{\infty\}, \oplus, \overline{0}, \otimes, \overline{1}),$$

with $\oplus = \min$, $\overline{0} = \infty$, $\otimes = +$, and $\overline{1} = 0$, to work with the negated logarithmic probabilities as weights, cf. Mohri et al. [2002]. Thus, decoding based on the Bayes' decision rule is also realized with weighted finite-state transducer operations on the

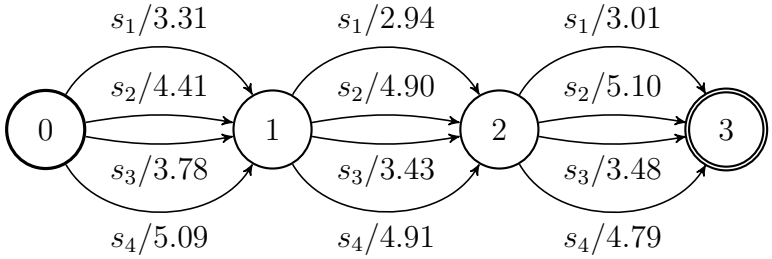


Figure A.4: Example for a weighted finite-state acceptor U that models the hidden Markov model acoustic state likelihood for a feature sequence of length 3 with 4 HMM acoustic states, cf. Povey et al. [2012].

tropical semiring. Since the logarithm function is strictly monotonically increasing, the Bayes' decision rule (Equation 2.4) can be formulated using logarithmic probabilities which leads to the *log-linear model*

$$\hat{\boldsymbol{w}} \approx \arg\max_{\boldsymbol{w}\in\mathcal{W}} \Big(\lambda \log(P(\boldsymbol{w})) + \max_{\boldsymbol{s}\in\mathcal{S}} (\log(P(\boldsymbol{X}|\boldsymbol{s})) + \log(P(\boldsymbol{s}|\boldsymbol{w}))) \Big). \tag{A.1}$$

In case negated logarithmic probabilities are used, we obtain

$$\hat{\boldsymbol{w}} \approx \arg\min_{\boldsymbol{w}\in\mathcal{W}} \Big(-\lambda \log(P(\boldsymbol{w})) + \min_{\boldsymbol{s}\in\mathcal{S}} (-\log(P(\boldsymbol{X}|\boldsymbol{s})) - \log(P(\boldsymbol{s}|\boldsymbol{w}))) \Big).$$

As can be easily seen, multiplication of probabilities is represented as the addition of (negated) logarithmic probabilities, and all operations in the latter log-linear model can be realized using the tropical semiring. Additionally, a *language model weight* (*LMWT*) λ is introduced to the acoustic and language model probabilities, cf. Yu and Deng [2015, p. 102]. The LMWT is a parameter that must be parameterized after training the model on development data to obtain accurate recognition results.

In order to describe the decoding process with finite-state transducers in Kaldi, let U be a weighted finite-state acceptor—a transducer with only a single transition symbol instead of input and output symbols—that models the acoustic model observation probabilities (or *acoustic likelihood*, as called in Kaldi) for a given sequence of features. Considering a feature sequence of length $T \in \mathbb{N}$, acceptor U has $T + 1$ states. Each state transition corresponds to the negated, logarithmic probability of observing the feature vector $\boldsymbol{x}_t$ at time step t being in state s_k. The acceptor is structured as exemplarily shown in Figure A.4.

The *search graph* S that is used for decoding, cf. Povey et al. [2012], is obtained by concatenating U with the input of the $HCLG$ transducer, i.e.,

$$S := U \circ HCLG.$$

Thus, the solution to the decoding problem formulation of Equation A.1 is equivalent to finding the best path through S, e.g., using the Viterbi search algorithm.

Some important implementation details were left out in the above description to illuminate the fundamental idea behind the decoding process with weighted finite-state transducers in Kaldi. These are, for example, pruning, determinization and minimization of transducers before concatenation, and the Kaldi lattice generation algorithm with separate storage of acoustic and graph costs. Detailed descriptions of the exact algorithms can be found in in [Mohri et al., 2002] on speech recognition with weighted finite-state transducers in general, in [Povey et al., 2012] on lattice generation in Kaldi, and in the official Kaldi documentation.

One major advantage of speech recognition with weighted finite-state transducers is that existing algorithms from toolboxes for mathematical operations on finite-state transducers, such as OpenFST by Allauzen et al. [2007], can be exploited for operations on the transducers. These algorithms are usually highly optimized to perform operations on FSTs. Moreover, these specialized algorithms reduce the redundancy and complexity of each speech recognition component stored as FSTs, which ultimately decreases the model size and increases inference efficiency for decoding. A further advantage is that the system is easily extensible with new model variants if the new variant can be expressed using finite-state transducers.

However, this also has the disadvantage of being less flexible due to being limited to operations and models that can be realized with FSTs. New algorithms or model variants must be able to be expressed as finite-state transducers. Specifics of algorithms in speech recognition are not necessarily considered for optimization, which might lead to slower performance or less precise results, due to approximation at certain steps, than an implementation taking these aspects into account, cf. for example, RASR by Rybach et al. [2009] which follows this approach.

A.2 The Fraunhofer IAIS Audio Mining System

A.2.1 Overview

The Fraunhofer IAIS Audio Mining system is designed to automatically create segmented and time-aligned transcriptions from long, unstructured audiovisual media files. Thus, the system combines automatic speech recognition with an audio analysis workflow, including segmentation of the audio signal, context detection, and speaker analysis using several pattern recognition algorithms.

Currently, the Fraunhofer IAIS Audio Mining system enables archivists, journalists, and hosts of audiovisual broadcast data to face the challenges caused by the continuously increasing amounts of long audiovisual recordings. This is achieved by making the files both text-searchable and structured. Thus, the amount of time a user needs to work with such 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 A.5. An embedded media player allows users to navigate to segments of specific speakers directly. A unique color represents different speakers in the time bar below the video. 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.

A.2.2 Audio Analysis in Audio Mining

In the following, we describe the audio analysis workflow as components of the Audio Mining system in more detail. This description is partly based on the work by Schmidt et al. [2016] from the Fraunhofer IAIS Institute. It is updated in this work to cover significant post-publication developments until 2022—particularly in the field of speech recognition. The schematic structure of the audio analysis workflow for one media file is illustrated in Figure A.6.

Audio Segmentation

The raw, unstructured audio signal is first cut into segments at speaker, channel, and environment changes by an audio segmentation algorithm proposed by Tritschler and Gopinath [1999]. For the segmentation, the *Bayesian information criterion* (BIC) is applied on full covariance Gaussian models of Mel-frequency cepstral coefficients.

Speech-/Non-Speech Detection

After segmentation, each segment is classified using a speech-/non-speech detection. Segments containing speech are passed to the following processing steps. The detection algorithm is a *Gaussian mixture model - universal background model* (GMM-UBM) approach trained for the classification task. In particular, the algorithm is trained to classify segments as *non-speech* on which no speech recognition is to be performed, such as music with vocals.

Figure A.5: Graphical web user interface of the Fraunhofer IAIS Audio Mining system in 2018. The video player on the left side uses the automatically generated transcript as subtitles. The colored time bar below the player displays segmentation and speaker clustering results to simplify navigation in large videos. Non-speech segments are represented by gray. A search mask in the upper right corner allows searching in a database of all indexed videos for words in the transcript, title, and other search modalities.

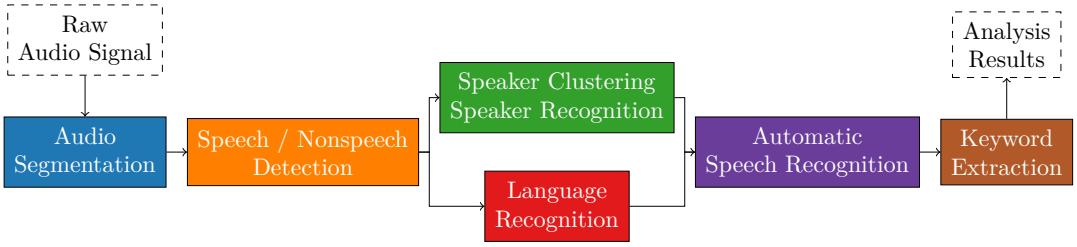


Figure A.6: Audio analysis workflow of the Audio Mining system, based on Schmidt et al. [2016] and updated to cover recent developments. The processing steps to obtain a structured transcription and searchable keywords for a raw audio signal are presented from left to right.

Speaker Clustering

This analysis step aims to identify all speech segments of the same speaker within one recording. As described above, only segments classified as speech in the previous step are considered for clustering.

We apply a BIC-based clustering algorithm in a bottom-up fashion, cf. Tritschler and Gopinath [1999]. By default, an attempt is made to automatically find the ideal number of speakers during clustering. Towards 2017, an additional option was added to Audio Mining to specify the maximum number of speakers in the given audio stream—in case they are known before analysis. Furthermore, the Fraunhofer IAIS performed experiments with i-vectors [Dehak et al., 2011] and added the possibility to perform speaker clustering with these features.

Speaker Recognition

Speaker recognition aims at finding all speech segments for known speakers in the given audio signal. Thus, Audio Mining users can search for well-known personalities, such as celebrities or politicians, in large collections of audiovisual media files. In Audio Mining, for example, users can search for quotes of individual politicians on specific topics by combining the search option for known speakers and transcription. The speaker recognition currently applied in Audio Mining is based on i-vectors and applies k-nearest neighbors classification.

The main difference between speaker clustering and recognition is that the speakers are not known for clustering, and segments with similar voices are grouped. In speaker recognition, however, the speakers are known, i.e., training samples exist for these speakers used for comparison. The speakers can be trained for individual use cases.

Language Recognition

Language recognition was introduced to Audio Mining in 2020 by Rieber [2020]. It enables automated detection of which language is spoken in the segment to select a speech recognizer with the appropriate language in the subsequent analysis. The language detection is based on the convolutional *Inception V3* model by Szegedy et al. [2016] that is well-known for image classification tasks. As input for the Inception V3, Rieber uses *Mel-scaled spectrograms* (that we referred to as *filter bank features* in the previous section) with a 128-dimensional feature space and fixed length of 10 seconds.

Currently, two different languages are available in Audio Mining to transcribe speech: German and English. Accordingly, the language recognition component is trained to detect three classes: *German*, *English*, and *Other*—for other languages

currently not supported by the system. However, more languages for Audio Mining are currently in preparation. For instance, French and Russian will become available in the near future.

Automatic Speech Recognition — Acoustic Model

Currently, Audio Mining supports transcription of German and English with several speech recognition models for different domains—and more languages are in preparation. However, at the beginning of the presented research work in 2017, only one German speech recognition model was used in Audio Mining. This ASR model is one of the models developed by Stadtschnitzer [2018] at the Fraunhofer IAIS institute in the years 2012–2018 and has been selected for release in Audio Mining. It is a cross-entropy trained hybrid DNN-HMM acoustic model, as described in Section 2.2, with a fully connected DNN, trained with the widely adopted Kaldi ASR toolkit by Povey et al. [2011]. It was trained on an in-house, 1000h large-scale, German broadcast corpus called *GerTV1000h* [Stadtschnitzer et al., 2014]. A detailed overview of this corpus is given in Section 3.4.

As Schmidt et al. [2016] state, CTC-RNN models trained with the EESEN-ASR-Toolkit [Miao et al., 2015a] have also been studied. In 2016–2017, these CTC-RNN models were considered for Audio Mining, as they provided slightly better results than the back than applied DNN-HMM model. However, this was abandoned, as further research, such as the work of Stadtschnitzer and the presented research work, showed that LF-MMI models, as described in Section 2.4.3, have the potential to achieve much better results for the German transcription use cases. The acoustic model currently used in Audio Mining is a robust acoustic model that has been trained as part of the presented work in the following Chapter 4.

Automatic Speech Recognition — Language Model

At the beginning of the presented research work, the language model used in Audio Mining was trained on broadcast text corpora consisting of 75 million words with a lexicon of about 500,000 words. This model is also used for most of the experiments presented in this work.

In recent years, the Fraunhofer IAIS has been training large-scale language models daily using recent crawls of German news sites. These models are deployed daily to the Audio Mining instances of clients—such as public broadcasters. This ensures that the Audio Mining system can recognize all new words and names entering the German language. These models are much larger than the previous language model and usually have a vocabulary of over 2 million words. One of these models trained on text data with 1.6 billion running words was also considered and studied for the oral history use case. We give a detailed overview of

the language models used for the experiments in the presented research work in Section 3.5.

Automatic Speech Recognition — Pronunciation Lexicon

The phonetic transcriptions for the lexicons of the different German language models are each obtained using a *grapheme-to-phoneme* (G2P) pronunciation model trained with Sequitur G2P [Bisani and Ney, 2008]. This model was trained using the German pronunciation database *Phonolex*[2] from the *Bavarian Archive for Speech Signals* (BAS), cf. Schiel [1998].

Not all G2P phonetizations are error-free since the pronunciations are generated automatically, learned from manually annotated German words. Wrong pronunciations in the lexicon lead to poor recognition of these words. Different work at the Fraunhofer IAIS studied approaches to improve further pronunciation modeling, such as Milde et al. [2017], who studied multi-task sequence-to-sequence models for the G2P task.

The G2P's susceptibility to errors is particularly strong for words that do not stem from the language the system was trained on—such as loanwords or anglicisms. This is due to their irregular pronunciation given the spelling compared to, for instance, native German words. In the context of the presented research work, student studies have been supervised that address this very problem and investigate different methods to improve the conversion for anglicisms: [Pritzen, Gref, Zühlke, and Schmidt, 2022] and [Pritzen, Gref, Schmidt, and Zühlke, 2021].

The work [Pritzen et al., 2022] proposes multi-task learning for sequence-to-sequence G2P systems where an additional anglicism classification task is added to a sequence-to-sequence G2P model. This approach aims to make the system aware of deviating pronunciation of anglicisms during training.

In [Pritzen et al., 2021], a *comparative pronunciation mapping* approach is proposed that compares likelihoods of German and English G2P systems to automatically detect the word heritage and select the appropriate G2P conversion. A phoneme mapping from the English to the German phoneme set was trained as part of the approach. The mapping uses German phoneme recognition applied on artificially created text-to-speech training samples of English words. This phoneme recognition is based on a German acoustic model trained in Chapter 4.

Keyword Extraction

In the last step of analysis in Audio Mining, keywords are extracted from the ASR generated transcript using a *term frequency - inverse document frequency* (*tf-idf*) approach. Audio Mining users can provide a blacklist or a whitelist to

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

ignore unwanted keywords in the keyword extraction and improve the client's application. The extracted keywords are stored in an external metadata format within an Audio Mining database, along with all results from the previous analysis steps. This allows users to search and filter media files that contain a specific topic.

Future Direction

The Audio Mining system is continuously improved to meet new client requirements and keep up with the state of art. For this purpose, new components are added, and existing components are enhanced regularly. For instance, the *speaker diarization* pipeline—the combination of audio segmentation and speaker clustering that identifies *who is speaking when*—is being replaced by more state-of-the-art approaches in two different student projects—both supervised as part of the presented research work. In particular, the system is being improved with regard to specific challenges such as fast speaker changes, short speaker segments, and double-talk.

Multi-modal emotion recognition and sentiment analysis have been developed since October 2020 in a research project focusing on the application for oral history interviews [Gref et al., 2022b]. This project is a collaboration of the *Haus der Geschichte* (HdG) foundation in Bonn and the Fraunhofer IAIS. For the multimodal recognition, the audio signal, the video stream, and the automatically generated ASR transcription are considered to recognize the emotions and the sentiment of the statements made in interviews. We plan to integrate this analysis component in Audio Mining or in a more general mining platform in the future to provide more sophisticated search options for audiovisual data.

Further work at the Fraunhofer IAIS aims to incorporate more languages into Audio Mining, on-the-fly lexicon extension, and improve computation time for analysis, for example, through parallelization or GPU-decoding for the speech analysis components.

A.2.3 Audio Mining Application for German Broadcasters

The Fraunhofer IAIS Audio Mining system has been continuously improved and developed further for more than ten years. The first and to date most significant use case is the application of Audio Mining at the archives of the *ARD*[3]—a joint organization of Germany's regional public-service broadcasters.[4]

Since 2015, Audio Mining has operated in the computing center of the ARD and is connected to the digital archives of the individual broadcasters. Nowadays,

[3] Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland; Working group of public broadcasters of the Federal Republic of Germany

[4] https://idw-online.de/de/news426319

the system automatically processes up to 2000 hours of audiovisual data per day.[5] The analysis results enable journalists and editors of the broadcasters to face the challenges of the ever-increasing amounts of audiovisual recordings in their archive. For example, Audio Mining enables the broadcasters to find interviews of different persons on specific topics or with certain quotations in the enormous audiovisual archives with little effort.

The Audio Mining system and its several audio analysis components were developed and optimized for precisely this application in the broadcasting sector. Therefore, annotated broadcast recordings and news texts are used for training the speech recognition components. From an audio analysis perspective, broadcast recordings are characterized by the recordings being recorded and post-processed using highly professional equipment. Thus, the recordings usually have excellent recording quality with well intelligible speech with almost no background noise, barely perceptible reverberation, and well-adjusted volume levels. The following section shows that most of this does not apply to the oral history interviews examined in the presented research work. Therefore, this mismatch between the broadcast training data and the oral history application poses major challenges to the Audio Mining speech recognition system.

[5] https://www.iais.fraunhofer.de/en/business-areas/speech-technologies/audio-mining-ard.html

B Appendix: Supplementary Results, Figures, and Tables

B.1 Automatic Speech Recognition Fundamentals

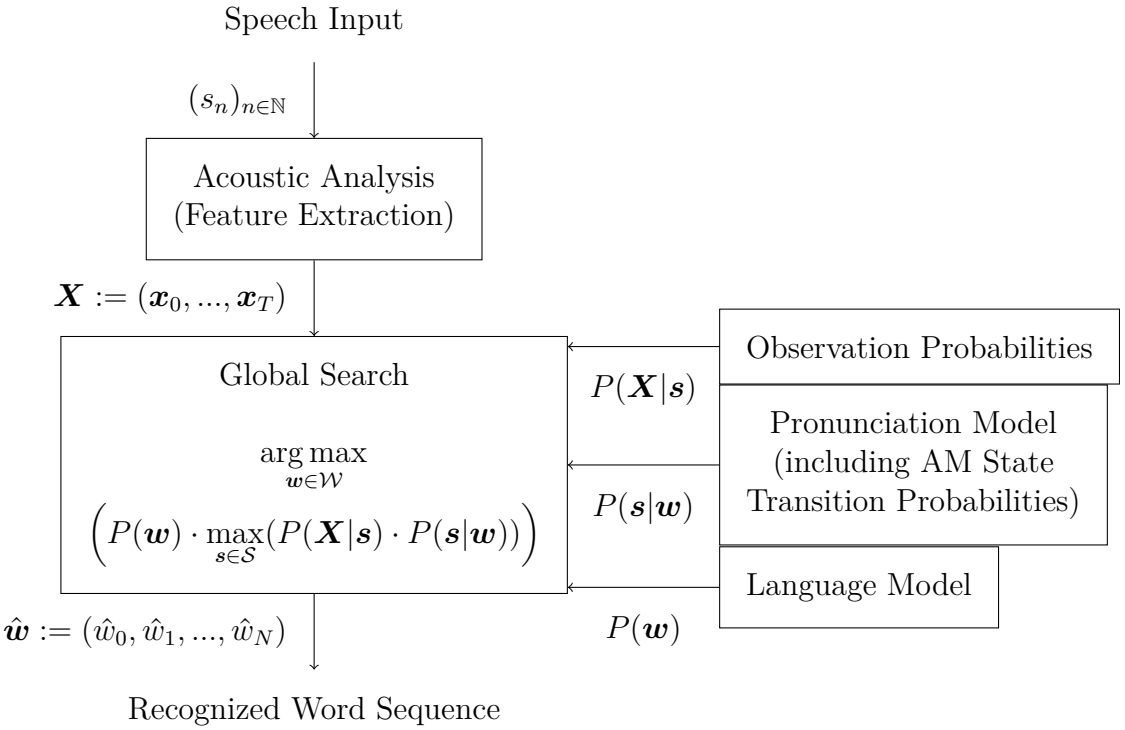


Figure B.1: Factorized Bayes' decision rule for hidden Markov model-based automatic speech recognition.

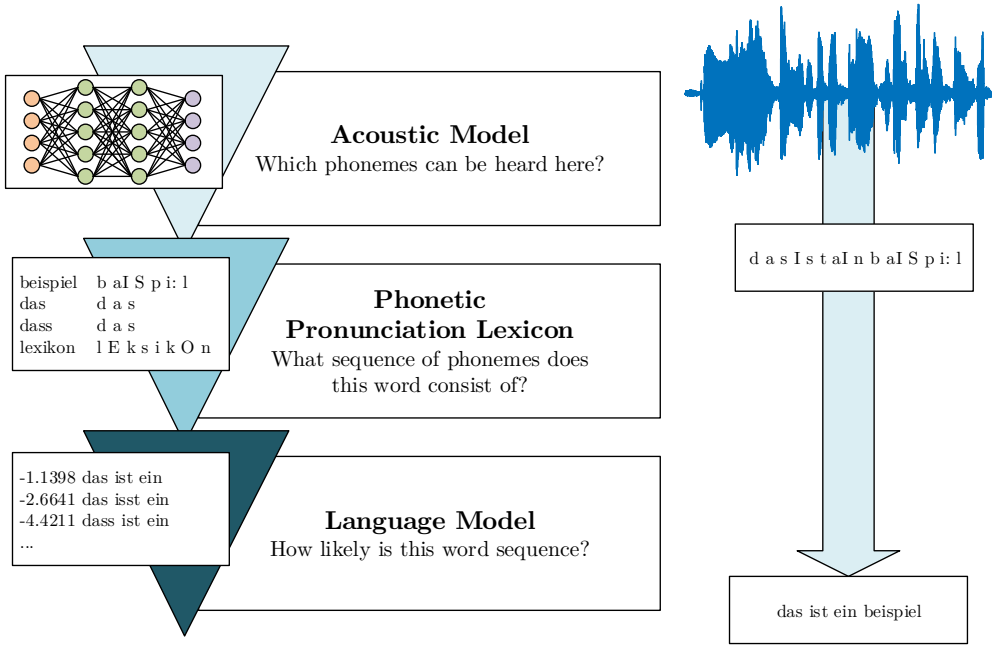


Figure B.2: Simplified, informal schematic structure of components in a large-vocabulary automatic speech recognition system.

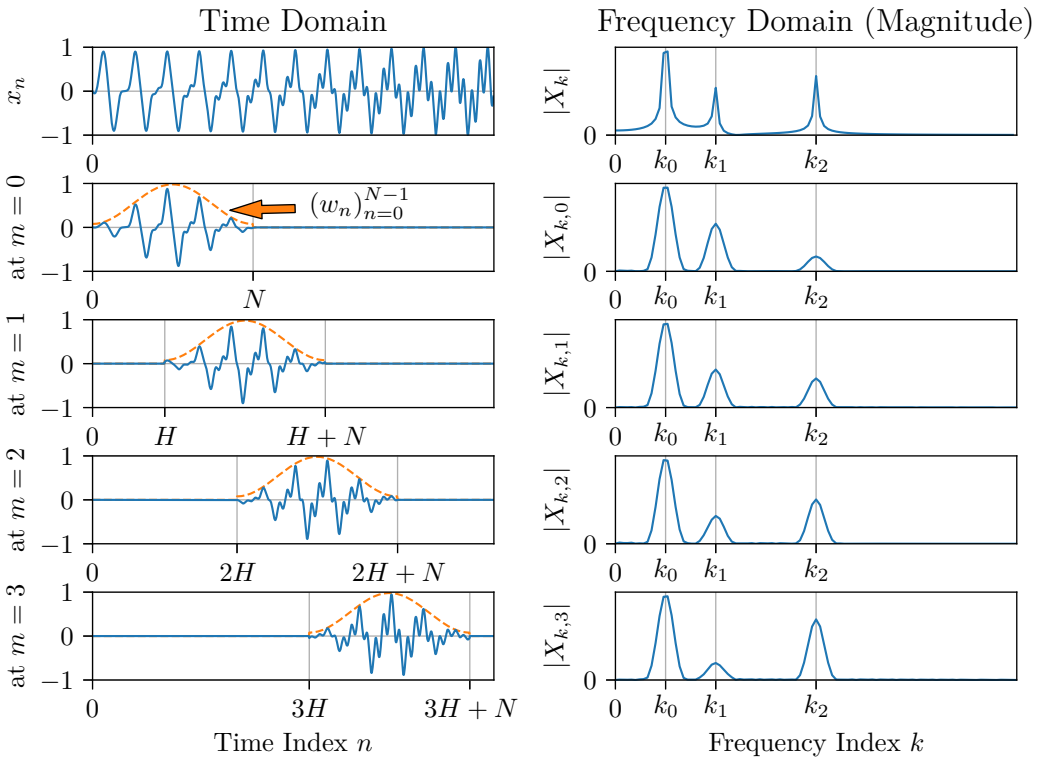


Figure B.3: Time-frequency analysis applying short-time Fourier transform on an exemplary harmonic signal with varying frequency characteristics along time. One can clearly distinguish the decreasing magnitude at frequency index k_1 and the increasing magnitude at frequency index k_2 along time from the magnitude spectrum using time-frequency analysis (frame index $m \in \{0, 1, 2, 3\}$).

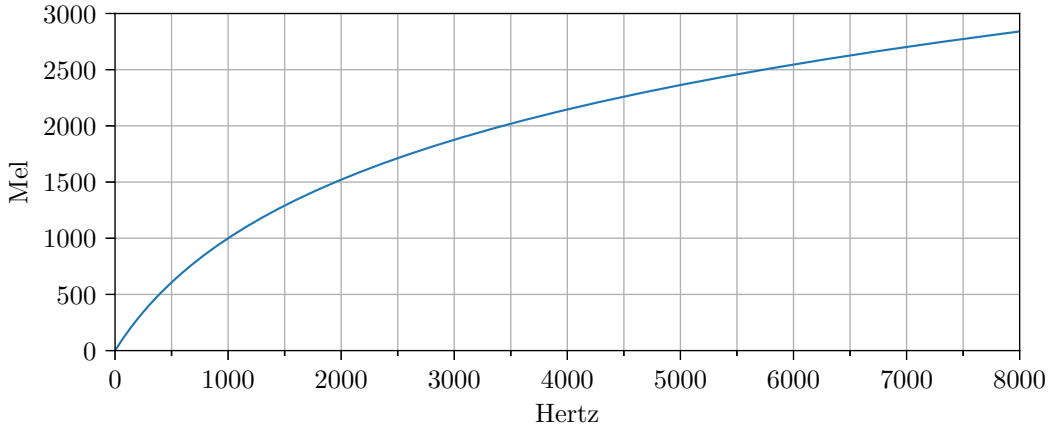


Figure B.4: Commonly used Mel scale (from 0 to 8000 Hz) maps Hertz frequencies to respective MEL frequencies.

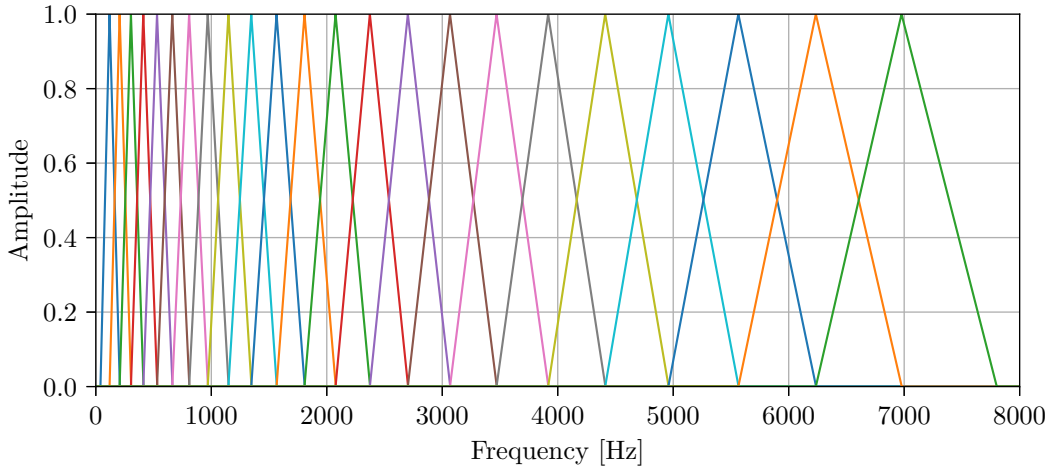


Figure B.5: Masks for Mel filter bank with 23 filters and cutoff frequencies at 40 Hz and 7800 Hz.

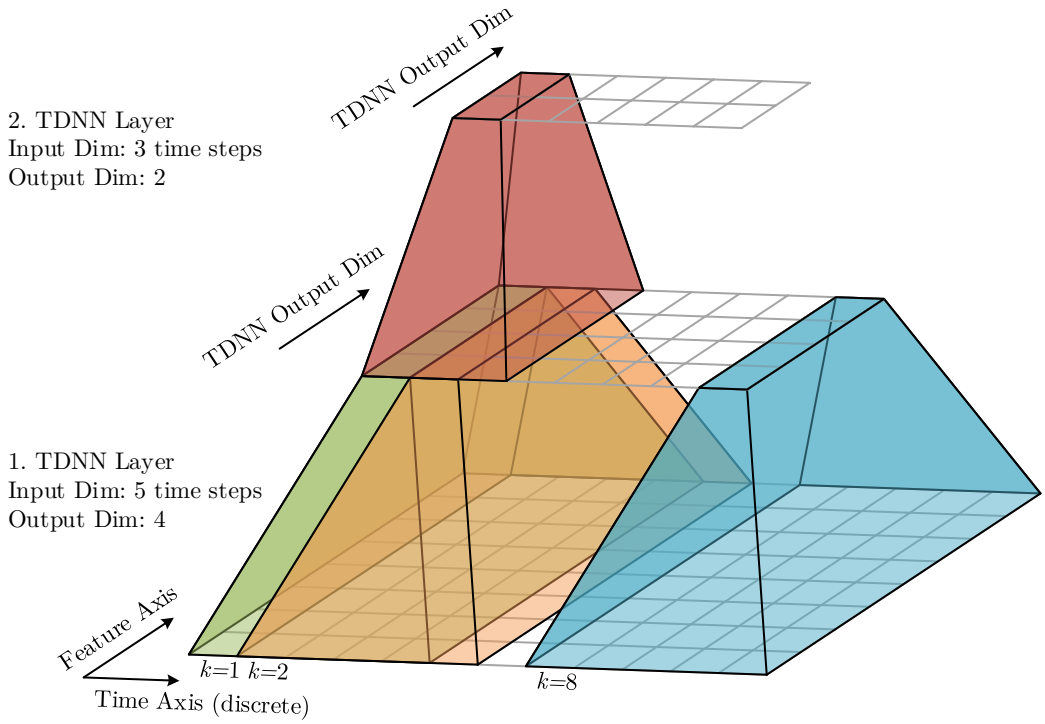


Figure B.6: Schematic architecture of a time delay neural network with two stacked TDNN layers. The semitransparent hexahedrons symbolize the sequentially applied layers along the time axis. For simplicity, the application of the layers is shown only for specific time steps: for the first TDNN layer at time steps one, two, and eight; for the second layer at time step one only.

B.2 Phone-Rate Estimation

Table B.1: Number of segments that could be aligned by the GMM-HMM acoustic model for phone rate estimation.

Data Set	Overall	Clean	3-fold	3-fold v2
GerTV1000h	773,631	768,684	770,081	771,089
DiSCo Planned Clean	1,364	1,362	1,362	1,362
DiSCo Planned Mix	2,200	2,197	2,200	2,200
DiSCo Spontaneous Clean	2,861	2,850	2,854	2,855
DiSCo Spontaneous Mix	1,650	1,619	1,640	1,644
German Broadcast 2016	227	222	222	223
Challenging Broadcast	593	553	563	568
Oral History	2,392	2,365	2,389	2,390
Interaction (Linguistics)	2,630	2,528	2,615	2,610
Spoken QALD-7	212	205	204	212

B.3 Two-Staged Acoustic Model Adaptation with Speaker-Aware Decoding

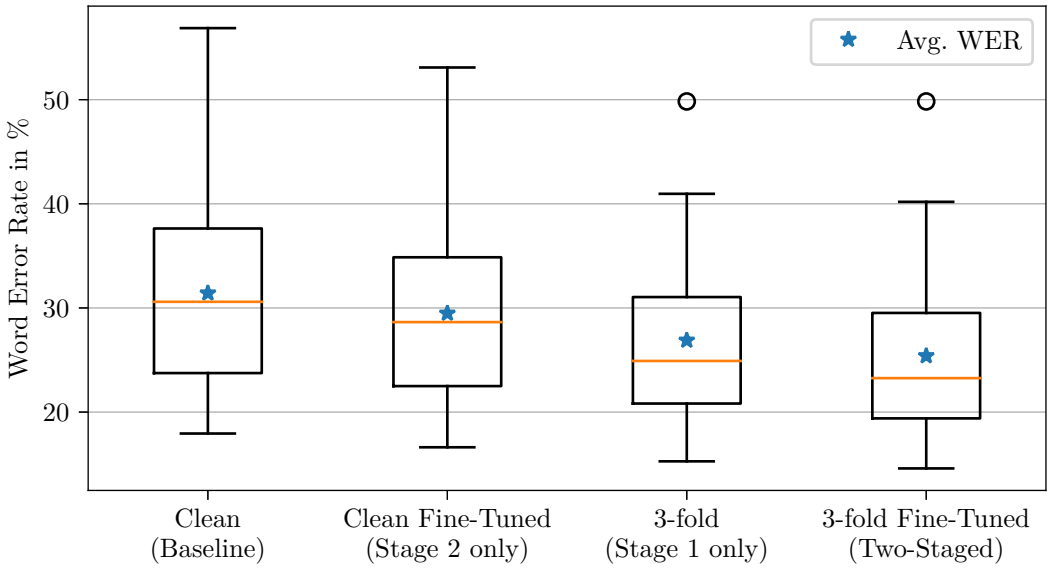


Figure B.7: Box plot diagram of the word error rates of the 35 interviews for each model in the leave-one-speaker-out experiments with speaker-aware decoding.

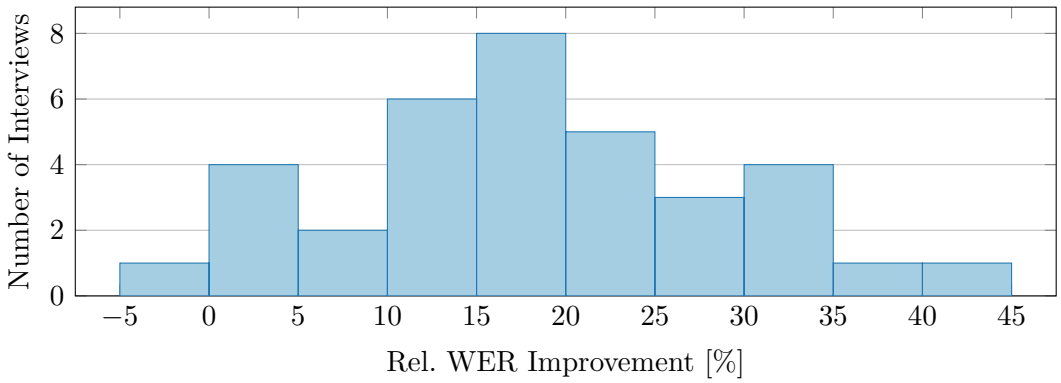


Figure B.8: Histogram of the relative word error rate improvements with the proposed approach two-staged acoustic model adaptation compared to the clean baseline for each leave-one-speaker-out experiment with speaker-aware decoding.

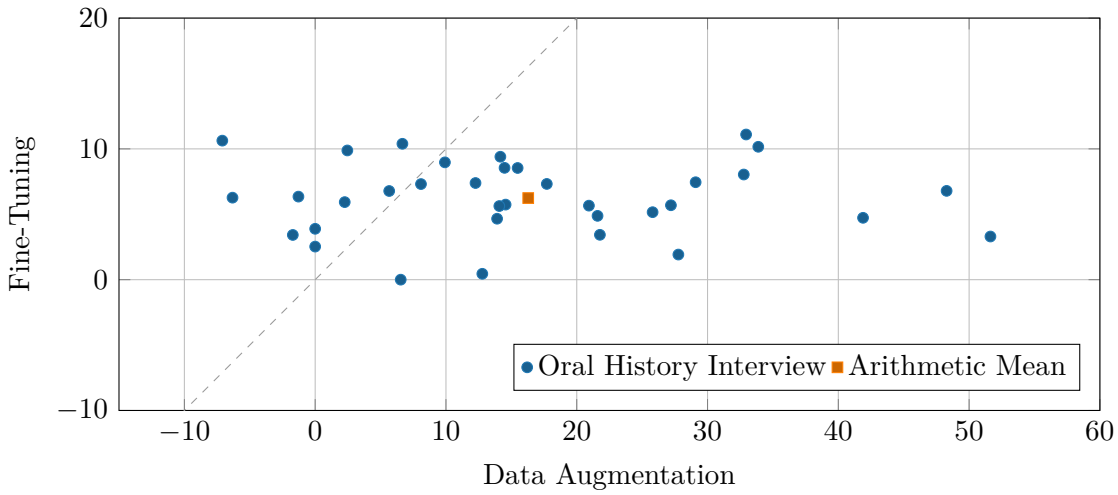


Figure B.9: Ablation study of the two-staged acoustic model adaptation with speaker-aware decoding by removing either data augmentation (Stage 1) or fine-tuning (Stage 2). Results are illustrated as a scatter plot of the relative word error rate increase compared to the proposed approach when one of the stages is removed. Positive values represent an increase in word error rate, i.e., the ASR performance deteriorates by removing this stage from the approach. The dashed diagonal axis marks the transition where both stages have an equal impact.

B.4 Acoustic Model Adaptation

Table B.2: Results of LF-MMI–TDNN-LSTM trained models (with per-frame dropout) using solely transcript-aligned oral history interviews. Training is performed from scratch.

	$\mathbf{OH_{49}}$	$\mathbf{OH_{99}}$	$\mathbf{OH_{150}}$	$\mathbf{OH_{150}^{10\%}}$
Training Set Size:	76 h	161 h	249 h	25 h
GerTV Dev Set	22.2	19.1	**17.9**	22.3
DiSCo Average	27.0	22.6	**21.0**	27.5
Planned Clean	18.3	16.3	**15.1**	18.9
Planned Mix	29.2	23.8	**21.3**	30.2
Spontaneous Clean	20.9	17.5	**16.2**	20.9
Spontaneous Mix	39.7	32.9	**31.4**	39.8
German Broadcast 2016	20.0	16.8	**15.9**	20.8
Challenging Broadcast	32.4	28.0	**26.3**	33.2
Oral History	35.3	30.5	**28.6**	37.3
Interaction	64.9	60.5	**58.6**	69.6
Spoken QALD-7	36.7	31.8	**28.4**	35.9

Table B.3: Detailed results for each HdG reference annotator of the comparison of adaptation experiments with reduced learning rate instead of reduced training data size. 1e-6/1e-7 is the default learning rate setup, 1e-7/1e-8 is the reduced learning rate.

Adaptation Data Set	$OH_{150}^{10\%}$	OH_{150}	OH_{150}
Adaptation Data Size	25 h	250 h	250 h
Learning Rate	1e-6/1e-7	1e-7/1e-8	1e-6/1e-7
(Mixed) Oral History	24.7	24.6	**23.9**
HdG Dev. Avg.	**16.7**±1.02	17.0±1.07	17.1±1.09
Transcriber A	**16.2**	16.4	16.4
Transcriber B	**16.1**	16.3	16.5
Transcriber C	**17.9**	18.2	18.4
HdG Test Avg.	**15.6**±0.33	15.9±0.32	16.1±0.36
Transcriber A	**15.3**	15.6	15.8
Transcriber B	**15.6**	15.9	16.1
Transcriber C	**16.0**	16.2	16.5

B.5 Multi-Staged Cross-Lingual Acoustic Model Adaptation

Table B.4: Detailed results for the DiSCo evaluation subsets of the multi-staged cross-lingual adaptation compared to two baselines and ablation studies. Results are reported both for the default and the large decoding language model as word error rates in percent.

	Baseline		**Ablation Study**			
	Broadcast 3-fold v1.1	**Oral History**	**Removing Stage 1**	**Removing Stage 2**	**Removing Stage 3**	**Proposed Approach**
Stage 1 (English)				×	×	×
Stage 2 (German Broadcast)	×		×		×	×
Stage 3 (Ger. Oral History)		×	×	×		×
DiSCo Average	**11.9**	27.9	12.4	20.8	**11.9**	12.4
+Large LM	12.2	28.7	12.5	20.7	**12.1**	12.4
Planned Clean	9.0	19.2	9.4	14.7	**8.8**	9.2
+Large LM	**11.1**	22.5	11.6	17.1	**11.1**	11.6
Planned Mix	11.0	31.1	11.1	21.6	**10.7**	**10.7**
+Large LM	**9.0**	19.0	9.3	14.5	9.2	9.3
Spontaneous Clean	**10.0**	21.2	10.9	16.6	10.3	11.1
+Large LM	9.9	30.4	10.0	20.2	**9.6**	9.9
Spontaneous Mix	**17.6**	40.2	18.4	30.3	**17.6**	18.4
+Large LM	**18.6**	43.0	19.0	31.1	**18.6**	18.9

C Appendix: Key Publications

In this appendix, the key publications of the present research work were attached for the review of the doctoral committee of the University of Bonn / Rheinische Friedrich-Wilhelms-Universität Bonn.

Some of the attached publications are protected by the copyright of the respective publisher and may not be published, copied, or distributed without their consent. A brief author contribution is given for each paper. Co-authors for whom no individual contribution is given contributed through scientific supervision.

C.1 Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research

> Michael Gref, Joachim Köhler, and Almut Leh. Improved transcription and indexing of oral history interviews for digital humanities research. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 3124–3131. European Language Resources Association (ELRA), 2018a. URL https://aclanthology.org/L18-1493

Author Contribution

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author. The co-authors of the publications contributed as follows:

Dr. Almut Leh described and discussed the oral history transcription use case for historical research and the oral history archive *Deutsches Gedächtnis* of the University of Hagen. In coordination with the thesis author, she also selected and provided the 35 oral history interviews for the oral history speech recognition test set introduced and investigated in the presented research work as the primary object of study, cf. summary in Section 3.4.6.

C.2 Improving Robust Speech Recognition for German Oral History Interviews Using Multi-Condition Training

Michael Gref, Christoph Schmidt, and Joachim Köhler. Improving robust speech recognition for German oral history interviews using multi-condition training. In *13th ITG Conference on Speech Communication*, pages 256–260. VDE / IEEE, 2018b. URL https://ieeexplore.ieee.org/document/8578034

Author Contribution

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author.

C.3 Two-Staged Acoustic Modeling Adaption for Robust Speech Recognition by the Example of German Oral History Interviews

Michael Gref, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Two-staged acoustic modeling adaption for robust speech recognition by the example of German oral history interviews. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 796–801, 2019. doi:10.1109/ICME.2019.00142

Author Contribution

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author.

C.4 Multi-Staged Cross-Lingual Acoustic Model Adaption for Robust Speech Recognition in Real-World Applications—A Case Study on German Oral History Interviews

Michael Gref, Oliver Walter, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Multi-staged cross-lingual acoustic model adaption for robust speech recognition in real-world applications - A case study on German oral history interviews. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6354–6362. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.780

Author Contribution

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author. The co-authors of the publications contributed as follows:

The training of the English model used as the source model for the cross-lingual adaptation, described in Section 6.4.1, was performed by Dr. Oliver Walter. He selected and prepared the English training data. Dr. Oliver Walter trained the English source model using the noise and reverberation data augmentation, model, and training routines proposed and studied by the thesis author. He also performed a preliminary evaluation of the English model using the two English test sets described in the aforementioned section.

C.5 Human and Automatic Speech Recognition Performance on German Oral History Interviews

> Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*, 2022b. URL https://arxiv.org/abs/2201.06841

Author Contribution

All presented approaches, experiments, findings, results, analyses, conclusions, figures, and texts are contributions of the thesis author. The co-authors of the publications contributed as follows:

Nike Matthiesen selected and provided the oral history interviews of the *Haus der Geschichte* (HdG) foundation. She coordinated and supervised the transcription of these interviews at the HdG in coordination with the thesis author. A summary of the HdG data is given by the thesis author in Section 3.4.10. The human word error rate investigations of the thesis author on this data are summarized in Section 3.6.

Bibliography

Hany Ahmed, Hazem Mamdouh, Salah Ashraf, Ali Ramadan, and Mohsen Rashwan. RDI-CU system for the 2019 arabic multi-genre broadcast challenge. In *ASRU Challenge Special Sessions: The 5th Edition of the Multi-Genre Broadcast Challenge: MGB-5*, 2019.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. The MGB-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019. doi:10.1109/asru46091.2019.9003960.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer Berlin Heidelberg, 2007. doi:10.1007/978-3-540-76336-9_3.

Tasos Anastasakos, John W. McDonough, Richard M. Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *4th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 1996. doi:10.1109/ICSLP.1996.607807.

Andrei Andrusenko, Aleksandr Laptev, and Ivan Medennikov. Towards a competitive end-to-end speech recognition for CHiME-6 dinner party transcription. In *21st Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2020. doi:10.21437/interspeech.2020-1074.

Paul Avan, Fabrice Giraudet, and Béla Büki. Importance of binaural hearing. *Audiology and Neurotology*, 20(Suppl. 1):3–6, 2015. doi:10.1159/000380741.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.0473.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016. doi:10.1109/icassp.2016.7472618.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983. doi:10.1109/TPAMI.1983.4767370.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers, 1986. doi:10.1109/icassp.1986.1169179.

James K. Baker. The DRAGON system–an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975. doi:10.1109/TASSP.1975.1162650.

Raimo Bakis. Continuous speech recognition via centisecond acoustic states. *The Journal of the Acoustical Society of America*, 59(S1):S97–S97, 1976. doi:10.1121/1.2003011.

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015. doi:10.1109/asru.2015.7404837.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1561–1565, 2018. doi:10.21437/Interspeech.2018-1768.

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 *7th International Conference on Language Resources and Evaluation (LREC)*, pages 1695–1699. European Language Resources Association (ELRA), 2010. URL https://aclanthology.org/L10-1244.

Leonard E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 1, 1972.

Leonard E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–364, 1967. doi:10.1090/s0002-9904-1967-11751-8.

Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6): 1554–1563, 1966. doi:10.1214/aoms/1177699147.

Leonard E. Baum and George Sell. Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227, 1968. doi:10.2140/pjm.1968.27.211.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. doi:10.1214/aoms/1177697196.

Peter Bell and Simon King. Diagonal priors for full covariance speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009. doi:10.1109/asru.2009.5373344.

Jacob Benesty, Man Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing*. Springer handbooks. Springer, 2008. ISBN 9783540491255. URL http://www.worldcat.org/oclc/612342075.

Klaus Beulen, Elmar Bransch, and Hermann Ney. State tying for context dependent phoneme models. In *5th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997. URL https://www.isca-speech.org/archive/eurospeech_1997/beulen97_eurospeech.html.

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008. doi:10.1016/j.specom.2008.01.002.

Hervé Bourlard and Nelson Morgan. Continuous speech recognition by connectionist statistical methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, 1993. doi:10.1109/72.286885.

Colin Breithaupt, Timo Gerkmann, and Rainer Martin. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *IEEE Signal Processing Letters*, 14(12):1036–1039, 2007. doi:10.1109/LSP.2007.906208.

Hennie Brugman and Albert Russel. Annotating multi-media/multi-modal resources with ELAN. In *4th International Conference on Language Resources and Evaluation (LREC)*, pages 2065–2068. European Language Resources Association (ELRA), 2004. URL https://aclanthology.org/L04-1285.

William Byrne, David S. Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas W. Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435, 2004. doi:10.1109/tsa.2004.828702.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus: A pre-announcement. pages 28–39. Springer Berlin Heidelberg, 2006. doi:10.1007/11677482_3.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016. doi:10.1109/icassp.2016.7472621.

William Chan, Daniel S. Park, Chris A. Lee, Yu Zhang, Quoc V. Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv:2104.02133 [cs.CL]*, abs/2104.02133, 2021. URL https://arxiv.org/abs/2104.02133.

Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur. Pronunciation and silence probability modeling for ASR. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 533–537, 2015. doi:10.21437/Interspeech.2015-198.

Kai Chen and Qiang Huo. Training deep bidirectional LSTM acoustic model for LVCSR by a context-sensitive-chunk BPTT approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1185–1193, 2016. doi:10.1109/taslp.2016.2539499.

Gaofeng Cheng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, and Yonghong Yan. An exploration of dropout with lstms. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1586–1590, 2017. doi:10.21437/Interspeech.2017-129.

Ekapol Chuangsuwanich, Yu Zhang, and James Glass. Multilingual data selection for training stacked bottleneck features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016. doi:10.1109/icassp.2016.7472711.

Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *4th International Conference on Language Resources and Evaluation (LREC)*, pages 69–71. European Language Resources Association (ELRA), 2004. URL https://aclanthology.org/L04-1500.

Tom Claes, Ioannis Dologlou, Louis ten Bosch, and Dirk Van Compernolle. A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6): 549–557, 1998. doi:10.1109/89.725321.

Jordan Cohen, Terri Kamm, and Andreas G. Andreou. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *The Journal of the Acoustical Society of America*, 97(5):3246–3247, 1995. doi:10.1121/1.411700.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv:1609.03193 [cs.LG]*, abs/1609.03193, 2016. URL https://arxiv.org/abs/1609.03193.

Ken H. Davis, R. Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952. doi:10.1121/1.1906946.

Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011. doi:10.1109/TASL.2010.2064307.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero. Recent advances in deep learning for speech research at microsoft. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. doi:10.1109/icassp.2013.6639345.

Lukas Drude, Jahn Heymann, Christoph Böddeker, and Reinhold Haeb-Umbach. NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing. In *13th ITG Conference on Speech Communication*, pages 216–220. VDE / IEEE, 2018. URL https://ieeexplore.ieee.org/document/8578026.

Jun Du, Tian Gao, Lei Sun, Feng Ma, Yi Fang, Di-Yuan Liu, Qiang Zhang, Xiang Zhang, Hai-Kun Wang, Jia Pan, Jian-Qing Gao, Chin-Hui Lee, and Jing-Dong Chen. The USTC-iFlytek systems for CHiME-5 challenge. In *5th International Workshop on Speech Processing in Everyday Environments (CHiME-5) Workshop*. ISCA, 2018. doi:10.21437/CHiME.2018-3.

Richard Dufour, Yannick Estève, and Paul Deléglise. Characterizing and detecting spontaneous speech: Application to speaker role recognition. *Speech Communication*, 56:1–18, 2014. doi:10.1016/j.specom.2013.07.007.

Ellen Eide and Herbert Gish. A parametric approach to vocal tract length normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996. doi:10.1109/icassp.1996.541103.

ETSI ES 201 108. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. Technical standard v1.1.3, European Telecommunications Standards Institute, 2003. URL https://portal.etsi.org/webapp/workprogram/Report_WorkItem.asp?wki_id=18820.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *9th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2008. doi:10.21437/interspeech.2008-271.

Siyuan Feng and Tan Lee. Improving cross-lingual knowledge transferability using multilingual TDNN-BLSTM with language-dependent pre-final layer. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018. doi:10.21437/interspeech.2018-1182.

Mark J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998. doi:10.1006/csla.1998.0043.

Mark J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3):272–281, 1999. doi:10.1109/89.759034.

Felix A. Gers and Jürgen Schmidhuber. Recurrent nets that time and count. In *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, pages 189–194, 2000. doi:10.1109/IJCNN.2000.861302.

Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000. doi:10.1162/089976600300015015.

Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur. Investigation of transfer learning for ASR using LF-MMI trained neural networks. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 279–286, 2017. doi:10.1109/ASRU.2017.8268947.

Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. Multilingual training of deep neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. doi:10.1109/icassp.2013.6639084.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992. doi:10.1109/icassp.1992.225858.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org.

Ramesh A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998. doi:10.1109/icassp.1998.675351.

Jan Gorisch, Michael Gref, and Thomas Schmidt. Using automatic speech recognition in spoken corpus curation. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6423–6428. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.790.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv:1211.3711 [cs.NE]*, 2012. URL https://arxiv.org/abs/1211.3711v1.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764–1772. PMLR, 2014. URL http://proceedings.mlr.press/v32/graves14.html.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, International Conference (ICML)*, pages 369–376, 2006. doi:10.1145/1143844.1143891.

Michael Gref, Joachim Köhler, and Almut Leh. Improved transcription and indexing of oral history interviews for digital humanities research. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 3124–3131. European Language Resources Association (ELRA), 2018a. URL https://aclanthology.org/L18-1493.

Michael Gref, Christoph Schmidt, and Joachim Köhler. Improving robust speech recognition for German oral history interviews using multi-condition training. In *13th ITG Conference on Speech Communication*, pages 256–260. VDE / IEEE, 2018b. URL https://ieeexplore.ieee.org/document/8578034.

Michael Gref, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Two-staged acoustic modeling adaption for robust speech recognition by the example of German oral history interviews. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 796–801, 2019. doi:10.1109/ICME.2019.00142.

Michael Gref, Oliver Walter, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Multi-staged cross-lingual acoustic model adaption for robust speech recognition in real-world applications - A case study on German oral history interviews. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 6354–6362. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.lrec-1.780.

Michael Gref, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke, and Joachim Köhler. A study on the ambiguity in human annotation of german oral history interviews for perceived emotion recognition and sentiment analysis. In *13th International Conference on Language Resources and Evaluation (LREC)*, pages 2022–2031. European Language Resources Association (ELRA), 2022a. URL https://aclanthology.org/2022.lrec-1.217.

Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler. Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*, 2022b. URL https://arxiv.org/abs/2201.06841.

Frantisek Grezl, Martin Karafiat, and Karel Vesely. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014. doi:10.1109/icassp.2014.6855089.

Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. End-to-end speech recognition using lattice-free MMI. In *19th Annual Conference of*

the International Speech Communication Association (Interspeech). ISCA, 2018. doi:10.21437/interspeech.2018-1423.

Reinhold Haeb-Umbach and Hermann Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992. doi:10.1109/icassp.1992.225984.

Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Jithendra Vepa, and Vincent Wan. The AMI meeting transcription system: Progress and performance. In *Machine Learning for Multimodal Interaction*, pages 419–431. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-69268-3. doi:10.1007/11965152_37.

Mary Harper. The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015. doi:10.1109/asru.2015.7404843.

William Hartmann, Tim Ng, Roger Hsiao, Stavros Tsakalidis, and Richard M. Schwartz. Two-stage data augmentation for low-resourced speech recognition. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2378–2382. ISCA, 2016. doi:10.21437/Interspeech.2016-1386.

Arjan van Hessen, Franciska de Jong, and Stef Scagliola. Der Einsatz von Sprachtechnologie in Oral-History-Sammlungen. In *Erinnern an Zwangsarbeit: Zeitzeugen-Interviews in der digitalen Welt*, pages 179–187. Metropol, 2013. ISBN 9783863311568. URL https://ris.utwente.nl/ws/portalfiles/portal/5577696/vanhessen-dejong-scagliola-vanhessen_ASR_2013.pdf.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. ISSN 1053-5888. doi:10.1109/MSP.2012.2205597.

Hans-Günter Hirsch. Extraction of robust features by combining noise reduction and fdlp for the recognition of noisy speech signals in hands-free mode. In *The REVERB Workshop (held in conjunction with ICASSP 2014 and HSCMA 2014)*, 2014.

Hans-Günter Hirsch and Michael Gref. On the influence of modifying magnitude and phase spectrum to enhance noisy speech signals. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1978–1982, 2017. doi:10.21437/Interspeech.2017-1173.

Hans-Günter Hirsch and Michael Gref. Keyword detection for the activation of speech assistants. In *13th ITG Conference on Speech Communication*, pages 186–190. VDE / IEEE, 2018. URL https://ieeexplore.ieee.org/document/8578020.

Hans-Günter Hirsch, Alexander Micheel, and Michael Gref. Keyword detection for the activation of speech dialogue systems. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung (ESSV)*, pages 2–9. TUDpress, Dresden, 2020. ISBN 978-3-959081-93-1. URL https://www.essv.de/paper.php?id=431.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. doi:10.1109/icassp.2013.6639081.

Xuedong Huang, James Baker, and Raj Reddy. A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103, 2014. doi:10.1145/2500887.

Mei-Yuh. Hwang and Xuedong Huang. Subphonetic modeling with markov states—senone. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992. doi:10.1109/icassp.1992.225979.

Hagen Jaeger, Michael Stadtschnitzer, Sofia B. Dias, and Leontios Hadjileontiadis. Development of a speech enhancement algorithm for the intervention of parkinson's disease within the i-PROGNOSIS framework. In *11th DPG Congress, Germany*. Zenodo, 2019. doi:10.5281/ZENODO.3695687.

Frederick Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976. doi:10.1109/proc.1976.10159.

Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *16th International Conference on Digital Signal Processing (DSP)*, pages 1–5, 2009. doi:10.1109/ICDSP.2009.5201259.

Ljubomir Josifovski. *Robust Automatic Speech Recognition with Missing and Unreliable Data*. PhD thesis, University of Sheffield, Department of Computer Science, UK, 2002.

Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, second edition, 2009. ISBN 9780135041963. URL https://www.worldcat.org/oclc/315913020.

Janez Kaiser, Bogomir Horvat, and Zdravko Kacic. A novel loss function for the overall risk criterion based discriminative training of HMM models. In *6th International Conference on Spoken Language Processing (ICSLP), 1st Annual Conference of the International Speech Communication Association (Interspeech)*, pages 887–890. ISCA, 2000. URL https://www.isca-speech.org/archive/icslp_2000/kaiser00_icslp.html.

Janez Kaiser, Bogomir Horvat, and Zdravko Kačič. Overall risk criterion estimation of hidden markov model parameters. *Speech Communication*, 38(3-4): 383–398, 2002. doi:10.1016/s0167-6393(02)00009-2.

Naoyuki Kanda, Rintaro Ikeshita, Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, Xiaofei Wang, Vimal Manohar, Nelson Enrique Yalta Soplin, Matthew Maciejewski, Szu-Jui Chen, Aswin Shanmugam Subramanian, Ruizhi Li, Zhiqi Wang, Jason Naradowsky, L. Paola Garcia-Perera, and Gregory Sell. The hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays. In *5th International Workshop on Speech Processing in Everyday Environments (CHiME-5) Workshop*. ISCA, 2018. doi:10.21437/CHiME.2018-2.

Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1435–1447, 2007a. doi:10.1109/tasl.2006.881693.

Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Speech and Audio Processing*, 15(4):1448–1460, 2007b. doi:10.1109/TASL.2007.894527.

Yuri Khokhlov, Alexander Zatvornitskiy, Ivan Medennikov, Ivan Sorokin, Tatiana Prisyach, Aleksei Romanenko, Anton Mitrofanov, Vladimir Bataev, Andrei Andrusenko, Mariya Korenevskaya, and Oleg Petrov. R-vectors: New technique for adaptation to room acoustics. In *20th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2019. doi:10.21437/interspeech.2019-2645.

Sameer Khurana, Ahmed Ali, and James Glass. Darts: Dialectal arabic transcription system. *arXiv:1909.12163v1 [cs.CL]*, 2019. URL https://arxiv.org/abs/1909.12163v1.

Do Yeong Kim, S. Umesh, M. J. F. Gales, Thomas Hain, and Philip C. Woodland. Using VTLN for broadcast news transcription. In *8th International Conference on Spoken Language Processing (ICSLP), 5th Annual Conference of the International Speech Communication Association (Interspeech)*, 2004. doi:10.21437/Interspeech.2004-191.

Brian Kingsbury. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009. doi:10.1109/icassp.2009.4960445.

Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1), 2016. doi:10.1186/s13634-016-0306-6.

Thomas Kisler, Uwe Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl, and Nina Pörner. BAS speech science web services - an update of current developments. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 3880–3885. European Language Resources Association (ELRA), 2016. URL https://aclanthology.org/L16-1614.

Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002. doi:10.1016/s0167-6393(01)00041-3.

Mendel Kleiner. *Electroacoustics*. CRC PressTaylor & Francis Group distributor, Abingdon Abingdon, 2013. ISBN 9781439836187.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3586–3589, 2015. doi:10.21437/Interspeech.2015-711.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for

robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5220–5224, 2017. doi:10.1109/ICASSP.2017.7953152.

Joachim Köhler, Michael Gref, and Almut Leh. KA3. Weiterentwicklung von Sprachtechnologien im Kontext der Oral History. *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, Schwerpunktheft: Digital Humanities und biographische Forschung*, 30(1-2/2017):44–59, 2019. doi:10.3224/bios.v30i1-2.05.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. doi:10.1109/icassp40776.2020.9053889.

Kshitiz Kumar, Chanwoo Kim, and Richard M. Stern. Delta-spectral cepstral coefficients for robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011. doi:10.1109/icassp.2011.5947425.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. Transfer learning for speech recognition on a budget. In *2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2017. doi:10.18653/v1/w17-2620.

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius - an open source real-time large vocabulary recognition engine. In *7th European Conference on Speech Communication and Technology (EUROSPEECH), 2nd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1691–1694, 2001. URL https://www.isca-speech.org/archive/eurospeech_2001/lee01c_eurospeech.html.

Almut Leh. Zeitzeugenkonserven. Interviews für nachfolgende Forschergenerationen im Archiv Deutsches Gedächtnis. *Archivar*, 71. Jg.(02):155–157, 2018. URL https://www.archive.nrw.de/sites/default/files/media/files/Archivar-2_2018.pdf.

Almut Leh and Doris Tausendfreund. Curation and dissemination of lifestory interviews for the humanities. In *2nd Conference on Biographical Data in a Digital World*, volume 2119 of *CEUR Workshop Proceedings*, pages 9–15. CEUR-WS.org, 2017. URL http://ceur-ws.org/Vol-2119/paper2.pdf.

Almut Leh, Joachim Köhler, Michael Gref, and Nikolaus Himmelmann. Speech analytics in research based on qualitative interviews. experiences from KA3. *VIEW Journal of European Television History and Culture*, 7(14):138–149, 2018. doi:10.18146/2213-0969.2018.jethc158.

Almut Leh, Michael Gref, and Joachim Köhler. Audio mining. advanced speech analytics for oral history. *Words and Silences/Palabras y Silencios*, (2018-2019):1–9, 2019. URL https://www.ioha.org/wp-content/uploads/2019/10/Leh_IOHA_2018_Audiomining_English.pdf.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965.

Stephen E. Levinson, Lawrence R. Rabiner, and Man Mohan Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62 (4):1035–1074, 1983. ISSN 0005-8580. doi:10.1002/j.1538-7305.1983.tb03114.x.

Bo Li, Tara N. Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean K. Chin, Khe Chai Sim, Ron J. Weiss, Kevin W. Wilson, Ehsan Variani, Chanwoo Kim, Olivier Siohan, Mitchel Weintraub, Erik McDermott, Richard Rose, and Matt Shannon. Acoustic modeling for google home. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 399–403, 2017. doi:10.21437/Interspeech.2017-234.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777, 2014. doi:10.1109/TASLP.2014.2304637.

Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, an imprint of Elsevier, Oxford, 2016. ISBN 9780128023983. doi:10.1016/C2014-0-02251-4.

Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. On the comparison of popular end-to-end models for large scale speech recognition. In *21st Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1–5. ISCA, 2020. doi:10.21437/Interspeech.2020-2846.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. Rethinking evaluation in ASR: Are our models robust enough? In *21st Annual Conference of*

the International Speech Communication Association (Interspeech). ISCA, 2021. doi:10.21437/interspeech.2021-1758.

Mike Lincoln, Iain McCowan, Jithendra Vepa, and Hari Krishna Maganti. The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 357–362, 2005. doi:10.1109/asru.2005.1566470.

Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997. doi:10.1016/s0167-6393(97)00021-6.

Richard P. Lippmann, Edward A. Martin, and Douglas B. Paul. Multi-style training for robust isolated-word speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 705–708, 1987. doi:10.1109/ICASSP.1987.1169544.

Jian Luo, Jianzong Wang, Ning Cheng, Edward Xiao, Jing Xiao, Georg Kucsko, Patrick O'Neill, Jagadeesh Balam, Slyne Deng, Adriana Flores, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, and Jason Li. Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2021. doi:10.1109/icme51207.2021.9428334.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015. doi:10.18653/v1/d15-1166.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. RWTH ASR systems for LibriSpeech: Hybrid vs attention. In *20th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2019. doi:10.21437/interspeech.2019-1780.

Jeff Ma, Francis Keith, Tim Ng, Man-Hung Siu, and Owen Kimball. Improving deliverable speech-to-text systems with multilingual knowledge transfer. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017. doi:10.21437/interspeech.2017-1058.

Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur. JHU kaldi system for arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017. doi:10.1109/asru.2017.8268956.

Iain Mccowan, J. Carletta, Wessel Kraaij, Simone Ashby, S. Bourban, M. Flynn, M. Guillemot, Thomas Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P. Wellner. The AMI meeting corpus. In *5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. EESEN: end-to-end speech recognition using deep RNN models and wfst-based decoding. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 167–174, 2015a. doi:10.1109/ASRU.2015.7404790.

Yajie Miao, Hao Zhang, and Florian Metze. Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1938–1949, 2015b. doi:10.1109/taslp.2015.2457612.

Péter Mihajlik, Tibor Fegyó, Bottyán Németh, Zoltán Tüske, and Viktor Trón. Towards automatic transcription of large spoken archives in agglutinating languages - hungarian ASR for the MALACH project. In *Text, Speech and Dialogue, 10th International Conference (TSD)*, pages 342–349, 2007. doi:10.1007/978-3-540-74628-7_45.

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Müllers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: IEEE Signal Processing Society Workshop*, 1999. doi:10.1109/nnsp.1999.788121.

Benjamin Milde, Christoph Schmidt, and Joachim Köhler. Multitask sequence-to-sequence models for grapheme-to-phoneme conversion. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017. doi:10.21437/interspeech.2017-1436.

Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88, 2002. ISSN 0885-2308. doi:10.1006/csla.2001.0184.

Nelson Morgan and Hervé Bourlard. Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE*, 83(5):742–772, 1995. doi:10.1109/5.381844.

Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Blind speech dereverberation with multi-channel linear

prediction based on short time fourier transform representation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008. doi:10.1109/icassp.2008.4517552.

Hermann Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999. doi:10.1109/79.790984.

Doug Oard. Can automatic speech recognition replace manual transcription? In *Oral History in the Digital Age*. Institute of Museum and Library Services, 2012.

Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. Prentice-Hall, Inc., Upper Saddle River, New Jersey, USA, second edition, 1999. ISBN 0-13-754920-2.

Andrew J. Oxenham. How we hear: The perception and neural coding of sound. *Annual Review of Psychology*, 69(1):27–50, 2018. doi:10.1146/annurev-psych-122216-011635. PMID: 29035691.

Cord Pagenstecher. Oral History und Digital Humanities. *BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, Schwerpunktheft: Digital Humanities und biographische Forschung*, Jg. 30, No. 1–2 (2017): 76–91, 2019a. doi:10.3224/bios.v30i1-2.07.

Cord Pagenstecher. Curating and analyzing oral history collections. *Selected papers from the CLARIN Annual Conference*, (159):144–151, 2019b. ISSN 1650-3740. URL https://ep.liu.se/ecp/article.asp?issue=159&article=015&volume=0.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015. doi:10.1109/icassp.2015.7178964.

Naveen Parihar and Joseph Picone. Aurora working group: DSR front end LVCSR evaluation AU/384/02. *Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University, technical report*, 2002.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *20th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2019. doi:10.21437/interspeech.2019-2680.

Douglas B. Paul and Janet M. Baker. The design for the wall street journal-based CSR corpus. In *2nd International Conference on Spoken Language Processing (ICSLP)*, 1992. doi:10.3115/1075527.1075614.

Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2440–2444. ISCA, 2015a. doi:10.21437/Interspeech.2015-527.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3214–3218, 2015b. doi:10.21437/Interspeech.2015-647.

Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur. Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters*, 25(3):373–377, 2018. doi:10.1109/lsp.2017.2723507.

Michael Picheny, Zoltán Tüske, Brian Kingsbury, Kartik Audhkhasi, Xiaodong Cui, and George Saon. Challenging the boundaries of speech recognition: The MALACH corpus. In *20th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2019. doi:10.21437/interspeech.2019-1907.

Daniel Povey and Brian Kingsbury. Evaluation of proposed modifications to MPE for large scale discriminative training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007. doi:10.1109/icassp.2007.366914.

Daniel Povey and Philip C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002. doi:10.1109/icassp.2002.5743665.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah. Boosted MMI for model and feature-space discriminative training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008. doi:10.1109/icassp.2008.4518545.

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 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011. IEEE Catalog No.: CFP11SRW-USB.

Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukas Burget, Arnab Ghoshal, Milos Janda, Martin Karafiat, Stefan Kombrink, Petr Motlicek, Yanmin Qian, Korbinian Riedhammer, Karel Vesely, and Ngoc Thang Vu. Generating exact lattices in the WFST framework. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012. doi:10.1109/icassp.2012.6288848.

Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. Parallel training of deep neural networks with natural gradient and parameter averaging. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL https://arxiv.org/abs/1410.7455.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2751–2755, 2016. doi:10.21437/Interspeech.2016-595.

Rohit Prabhavalkar, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017. doi:10.21437/interspeech.2017-233.

Julia Pritzen, Michael Gref, Christoph Schmidt, and Dietlind Zühlke. A comparative pronunciation mapping approach using G2P conversion for anglicisms in German speech recognition. In *14th ITG Conference on Speech Communication*, pages 24–28. VDE / IEEE, 2021. URL https://ieeexplore.ieee.org/document/9657500.

Julia Pritzen, Michael Gref, Dietlind Zühlke, and Christoph Andreas Schmidt. Multitask learning for grapheme-to-phoneme conversion of anglicisms in German speech recognition. In *13th International Conference on Language Resources and Evaluation (LREC)*, pages 3242–3249. European Language Resources Association (ELRA), 2022. URL https://aclanthology.org/2022.lrec-1.346.

Josef Psutka, Pavel Ircing, Josef V. Psutka, Vlasta Radová, William J. Byrne, Jan Hajic, Samuel Gustman, and Bhuvana Ramabhadran. Automatic transcription of czech language oral history in the MALACH project: Resources and initial

experiments. In *Text, Speech and Dialogue, 5th International Conference (TSD)*, pages 253–260, 2002. doi:10.1007/3-540-46154-X_34.

Josef Psutka, Pavel Ircing, Josef V. Psutka, Jan Hajic, William J. Byrne, and Jirí Mírovský. Automatic transcription of czech, russian, and slovak spontaneous speech in the MALACH project. In *9th European Conference on Speech Communication and Technology (EUROSPEECH), 6th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1349–1352, 2005. URL https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2005/i05_1349.pdf.

Bhargav Pulugundla, Murali Karthick Baskar, Santosh Kesiraju, Ekaterina Egorova, Martin Karafiát, Lukáš Burget, and Jan Černocký. BUT system for low resource indian language ASR. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018. doi:10.21437/interspeech.2018-1302.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi:10.1109/5.18626.

Lawrence R. Rabiner and Stephen E. Levinson. A speaker-independent, syntax-directed, connected word recognition system based on hidden markov models and level building. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):561–573, 1985. doi:10.1109/TASSP.1985.1164586.

Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. Probing the information encoded in x-vectors. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019. doi:10.1109/asru46091.2019.9003979.

Bhuvana Ramabhadran, Jing Huang, and Michael Picheny. Towards automatic transcription of large spoken archives - english ASR for the MALACH project. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003. doi:10.1109/icassp.2003.1198756.

Shakti P. Rath, Daniel Povey, Karel Veselý, and Jan Cernocký. Improved feature processing for deep neural networks. In *14th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 109–113. ISCA, 2013. doi:10.21437/Interspeech.2013-48.

Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeffrey Hetherly, Cory Stephenson, and Karl

Ni. Voices obscured in complex environmental settings (VOiCES) corpus. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018. doi:10.21437/interspeech.2018-1454.

Joscha Simon Rieber. Spoken language recognition using convolutional neural networks. Technical report, 2020. URL https://pub.towardsai.net/spoken-language-recognition-using-convolutional-neural-networks-6aec5963eb18.

Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. WSJCAM0: a british english speech corpus for large vocabulary continuous speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 81–84, 1995. doi:10.1109/ICASSP.1995.479278.

David Rybach, Christian Gollan, Georg Heigold, Björn Hoffmeister, Jonas Lööf, Ralf Schlüter, and Hermann Ney. The RWTH aachen university open source speech recognition system. In *10th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2111–2114, 2009. doi:10.21437/Interspeech.2009-604.

Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *15th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 338–342, 2014. doi:10.21437/Interspeech.2014-80.

Elizabeth Salesky, Jessica Ray, and Wade Shen. Operational assessment of keyword search on oral history. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 317–321. European Language Resources Association (ELRA), 2016. URL https://aclanthology.org/L16-1049.

George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013. doi:10.1109/asru.2013.6707705.

George Saon, Hong-Kwang Jeff Kuo, Steven J. Rennie, and Michael Picheny. The IBM 2015 english conversational telephone speech recognition system. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3140–3144, 2015. doi:10.21437/Interspeech.2015-632.

George Saon, Tom Sercu, Steven J. Rennie, and Hong-Kwang Jeff Kuo. The IBM 2016 english conversational telephone speech recognition system. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 7–11, 2016. doi:10.21437/Interspeech.2016-1460.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. English conversational telephone speech recognition by humans and machines. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 132–136, 2017. doi:10.21437/Interspeech.2017-405.

Florian Schiel. Speech and speech releated resources at BAS. In *1st International Conference on Language Resources and Evaluation (LREC)*, pages 343–350. European Language Resources Association (ELRA), 1998.

Christoph Schmidt, Michael Stadtschnitzer, and Joachim Köhler. The Fraunhofer IAIS audio mining system: Current state and future directions. In *12th ITG Conference on Speech Communication*, pages 115–119. VDE / IEEE, 2016. URL http://ieeexplore.ieee.org/document/7776158.

Armin Sehr, Christian Hofmann, Roland Maas, and Walter Kellermann. Multi-style training of HMMS with stereo data for reverberation-robust speech recognition. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. IEEE, 2011. doi:10.1109/hscma.2011.5942396.

Xian Shi, Qiangze Feng, and Lei Xie. The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results. *arXiv:2007.05916v1 [eess.AS]*, 2020. URL https://arxiv.org/abs/2007.05916v1.

Olivier Siohan, Bhuvana Ramabhadran, and Geoffrey Zweig. Speech recognition error analysis on the english MALACH corpus. In *8th International Conference on Spoken Language Processing (ICSLP), 5th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2004. doi:10.21437/Interspeech.2004-171.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018. doi:10.1109/icassp.2018.8461375.

Hagen Soltau, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon, and Geoffrey Zweig. The IBM 2004 conversational telephony system for rich transcription. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 205–208, 2005. doi:10.1109/ICASSP.2005.1415086.

Hagen Soltau, Hank Liao, and Hasim Sak. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3707–3711, 2017. doi:10.21437/Interspeech.2017-1566.

Michael Stadtschnitzer. *Robust Speech Recognition for German and Dialectal Broadcast Programmes*. PhD thesis, University of Bonn, Germany, 2018. URL https://hdl.handle.net/20.500.11811/7658.

Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein, and Joachim Köhler. Exploiting the large-scale german broadcast corpus to boost the fraunhofer IAIS speech recognition system. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 3887–3890. European Language Resources Association (ELRA), 2014. URL https://aclanthology.org/L14-1664.

Stanley Smith Stevens, John E. Volkmann, and Edwin B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. doi:10.1121/1.1915893.

Andreas Stolcke and Jasha Droppo. Comparing human and machine errors in conversational speech transcription. In *18th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 137–141, 2017. doi:10.21437/Interspeech.2017-1544.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi:10.1109/cvpr.2016.308.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.findings-emnlp.295.

Hao Tang, Wei-Ning Hsu, François Grondin, and James Glass. A study of enhancement, augmentation and autoencoder methods for domain adaptation in distant speech recognition. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018. doi:10.21437/interspeech.2018-2030.

Sibo Tong, Philip N. Garner, and Hervé Bourlard. Cross-lingual adaptation of a CTC-based multilingual acoustic model. *Speech Communication*, 104:39–46, 2018. doi:10.1016/j.specom.2018.09.001.

Sibo Tong, Philip N. Garner, and Herve Bourlard. An investigation of multilingual ASR using end-to-end LF-MMI. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019. doi:10.1109/icassp.2019.8683338.

Edmondo Trentin and Marco Gori. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 2001. doi:10.1016/S0925-2312(00)00308-8.

Alain Tritschler and Ramesh A. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999. URL https://www.isca-speech.org/archive_v0/eurospeech_1999/e99_0679.html.

Alexandros Tsilfidis, Adam Westermann, Jörg M. Buchholz, Eleftheria Georganti, and John Mourjopoulos. Binaural dereverberation. In *The Technology of Binaural Listening*, pages 359–396. Springer Berlin Heidelberg, 2013. doi:10.1007/978-3-642-37762-4_14.

Juliane Turzynski. Zeitliche Alignierung von Audiodaten und Transkripten mit Abweichungen. Master's thesis, TH Köln – University of Applied Sciences, 2017. URL http://publica.fraunhofer.de/dokumente/N-474180.html.

Sei Ueno, Takafumi Moriya, Masato Mimura, Shinsuke Sakai, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono, and Tatsuya Kawahara. Encoder transfer for attention-based acoustic-to-word speech recognition. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2424–2428, 2018. doi:10.21437/Interspeech.2018-1424.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 7th open challenge on question answering over linked data (QALD-7). In *Semantic Web Challenges - 4th*

SemWebEval Challenge at (ESWC), pages 59–69, 2017. doi:10.1007/978-3-319-69146-6_6.

Daniel Vasquez, Rainer Gruhn, and Wolfgang Minker. *Hierarchical Neural Network Structures for Phoneme Recognition*. Springer-Verlag GmbH, 2012. ISBN 3642344259. doi:10.1007/978-3-642-34425-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762v5 [cs.CL]*, 2017. URL https://arxiv.org/abs/1706.03762v5.

Keith Vertanen. An overview of discriminative training for speech recognition. Technical report, 2005.

Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *14th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2345–2349. ISCA, 2013. doi:10.21437/Interspeech.2013-548.

Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1–3): 133–147, 1998. ISSN 0167-6393. doi:10.1016/S0167-6393(98)00033-8.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Comput. Speech Lang.*, 46:535–557, 2017. doi:10.1016/j.csl.2016.11.005.

Paul Voigtlaender, Patrick Doetsch, Simon Wiesler, Ralf Schluter, and Hermann Ney. Sequence-discriminative training of recurrent neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015. doi:10.1109/icassp.2015.7178341.

Apoorv Vyas, Srikanth Madikeri, and Hervé Bourlard. Comparing CTC and LFMMI for out-of-domain adaptation of wav2vec 2.0 acoustic model. In *22nd Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2021. doi:10.21437/interspeech.2021-1683.

Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical report, USA, 2004.

Dong Wang and Thomas Fang Zheng. Transfer learning for speech and language processing. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237, 2015. doi:10.1109/APSIPA.2015.7415532.

Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018, 2019. doi:10.3390/sym11081018.

Guangsen Wang and Khe Chai Sim. Sequential classification criteria for nns in automatic speech recognition. In *12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 441–444, 2011. doi:10.21437/Interspeech.2011-170.

Yao Wang, Michael Gref, Oliver Walter, and Christoph Schmidt. Bilingual i-vector extractor for DNN hybrid acoustic model training in German speech recognition systems. In *14th ITG Conference on Speech Communication*, pages 29–33. VDE / IEEE, 2021. URL https://ieeexplore.ieee.org/document/9657501.

Wayne Ward. Understanding spontaneous speech. In *Workshop on Speech and Natural Language (HLT)*. Association for Computational Linguistics, 1989. doi:10.3115/100964.100975.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017. doi:10.1109/jstsp.2017.2763455.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423, 2017a. doi:10.1109/taslp.2017.2756440.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. The microsoft 2016 conversational speech recognition system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5255–5259, 2017b. doi:10.1109/ICASSP.2017.7953159.

Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5934–5938, 2018. doi:10.1109/ICASSP.2018.8461870.

Haihua Xu, Hang Su, Chongjia Ni, Xiong Xiao, Hao Huang, Eng Siong Chng, and Haizhou Li. Semi-supervised and cross-lingual knowledge transfer learnings for DNN hybrid acoustic models under low-resource conditions. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016. doi:10.21437/interspeech.2016-1099.

Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, Lirong Dai, and Qingfeng Liu. Fast adaptation of deep neural network based on discriminant codes for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1713–1725, 2014. doi:10.1109/taslp.2014.2346313.

Takuya Yoshioka, Anton Ragni, and Mark J. F. Gales. Investigation of unsupervised adaptation of DNN acoustic models with filter bank input. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014. doi:10.1109/icassp.2014.6854825.

Steve Young, Gunnar Evermann, M. J. F. Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, James Odell, Dave Ollason, Daniel Povey, Anton Ragni, Valtcho Valtchev, Philip C. Woodland, and Chao Zhang. *The HTK Book (version 3.5a)*. 2015.

Dong Yu and Li Deng. *Automatic speech recognition: A deep learning approach*. Signals and Communication Technology. Springer, London, 2015. ISBN 978-1-4471-5778-6.

Zbynek Zajic, Lucie Skorkovska, Petr Neduchal, Pavel Ircing, Josef V. Psutka, Marek Hruz, Ales Prazak, Daniel Soutner, Jan Švec, Lukas Bures, and Ludek Muller. Towards processing of the oral history interviews and related printed documents. In *11th International Conference on Language Resources and Evaluation (LREC)*, pages 2099–2104. European Language Resources Association (ELRA), 2018. ISBN 979-10-95546-00-9. URL https://aclanthology.org/L18-1331.

Albert Zeyer, Ralf Schlüter, and Hermann Ney. Towards online-recognition with deep bidirectional LSTM acoustic models. In *17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016. doi:10.21437/interspeech.2016-759.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. Improved training of end-to-end attention models for speech recognition. In *19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018. doi:10.21437/interspeech.2018-1616.

Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yao, Sanjeev Khudanpur, and James Glass. Highway long short-term memory RNNS for distant speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016. doi:10.1109/icassp.2016.7472780.