universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Domain Adaptation of Whisper Models to German Medical Automatic Speech Recognition

verfasst von | submitted by

Valentina Hofecker BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 587
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Multilingual Technologies
Betreut von | Supervisor:	Miguel Angel Rios Gaona PhD

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor, Mr. Angel Rios Gaona, PhD, for his extensive and ongoing support. I am more than grateful for all his efforts and highly appreciate his calm manner, which brought me back down to earth in stressful moments. I would like to extend my sincere thanks to Bettina Müller at 4voice AG and Christof Strauß at Speech Processing Solutions GmbH for their contributions to the recording of the audios. Additionally, I am thankful to Martin Müller from 4voice AG for providing the data for the creation of our datasets. I also want to say a big *Thank You* to my boyfriend, Philipp, for always having an open ear. His endless support and encouragement were extremely helpful during the more challenging moments of writing this thesis. Finally, I am profoundly grateful for all my colleagues at the MLT24 Master's program. Your support and friendship means the world to me and I am extremely glad we got to experience these last two years together.

Abstract

Despite the advancements of Automatic Speech Recognition (ASR) technologies in healthcare settings, the systems face difficulties in accurately recognizing domain-specific, medical terminology. This challenge is particularly important in the German-speaking medical domain, where transcription quality is comparatively poorer than in general domains. There has been extensive research on creating multilingual ASR datasets to refine ASR technologies. However, the German medical sector, especially with Austrian Standard German pronunciation, remains significantly underrepresented. Previous studies have broadly explored domain adaptation to transfer knowledge to the German domain, but not specifically for the medical sector or for Austrian Standard German pronunciation. In this thesis, we address these gaps and challenges by applying domain adaptation and finetuning approaches on Whisper models, SOTA transformer-based ASR models, proposed by OpenAI. Initially, we conduct a benchmarking project to evaluate the baseline performance of the original Whisper models to provide a clear reference point for subsequent improvements. For this, we create our own custom ASR dataset, MedASR-DE-AT, containing 11.5 hours of German medical speech, specifically with Austrian Standard German pronunciation, accompanied by transcription data. Furthermore, we build an additional dataset solely for testing, MedASR-DE-test, containing Standard German pronunciation to explore the impact of dialectal variations on model performance. We finetune the Whisper models using the MedASR-DE-AT dataset to enhance transcription quality. Through this process, we achieve a substantial improvement, reducing the Word Error Rate (WER) for the Whisper-medium model by 14.52%. Additionally, we perform error annotation on the original and finetuned Whisper models and present a thorough error analysis. This work contributes to the enhancement of ASR technology's accuracy and applicability in the medical domain, especially in the German-speaking sector.

Zusammenfassung

Trotz der weit verbreiteten Nutzung von Systemen zur automatischen Spracherkennung (ASR) im Gesundheitswesen bleibt die automatische Transkription von domänenspezifischen medizinischen Texten eine erhebliche Herausforderung. Diese Herausforderung ist besonders im deutschsprachigen medizinischen Bereich ausgeprägt, wo die Transkriptionsqualität schlechter ist als in allgemeinen Bereichen. Umfangreiche Forschungsarbeiten zur Erstellung mehrsprachiger ASR-Datensätze haben bereits zur Verfeinerung von ASR-Technologien beigetragen. Der deutsche medizinische Bereich, insbesondere mit österreichischer Standardaussprache, ist jedoch weiterhin deutlich unterrepräsentiert. Frühere Studien konzentrierten sich auf die Domänenanpassung zur Übertragung von Wissen auf die deutsche Domäne im Allgemeinen, jedoch nicht speziell für den medizinischen Sektor oder die österreichische Standardaussprache. Diese Arbeit adressiert die genannten Lücken und Herausforderungen durch die Implementierung von Domänenanpassungs- und Feinabstimmungsstrategien auf den Whisper-Modellen von OpenAI, welche zu den modernsten Transformer-basierten ASR-Modellen zählen. Zunächst führen wir ein Benchmarking-Projekt durch, um die Ausgangsleistung der ursprünglichen Whisper-Modelle zu bewerten und einen klaren Bezugspunkt für spätere Verbesserungen zu schaffen. Zu diesem Zweck erstellen wir einen eigenen ASR-Datensatz, MedASR-DE-AT, der 11,5 Stunden deutscher medizinischer Sprache enthält, insbesondere mit österreichischer Standardaussprache, begleitet von Transkriptionsdaten. Darüber hinaus erstellen wir einen zusätzlichen Datensatz, MedASR-DE-test, ausschließlich zu Testzwecken, der die standarddeutsche Aussprache enthält, um die Auswirkungen dialektaler Unterschiede auf die Modellleistung zu untersuchen. Wir trainieren die Whisper-Modelle anhand des MedASR-DE-AT-Datensatzes und führen Experimente zur Feinabstimmung durch, um die Transkriptionsqualität weiter zu verbessern. Durch diesen Prozess können wir die Wortfehlerrate (WER) des Whisper-Medium-Modells um 14.52% senken. Außerdem führen wir eine Fehlerannotation für das ursprüngliche und das feinabgestimmte Whisper-Modell durch und präsentieren eine umfassende Fehleranalyse. Somit trägt diese Arbeit zur Verbesserung der Genauigkeit und Anwendbarkeit von ASR-Technologien im medizinischen Bereich bei, insbesondere im deutschsprachigen Raum.

Contents

List of Tables **13**

List of Figures **15**

1 Introduction **1**
- 1.1 Challenges in Automatic Speech Recognition . . . 3
- 1.2 Medical Automatic Speech Recognition . . . 4
  - 1.2.1 Challenges in Medical Automatic Speech Recognition . . . 4
- 1.3 German Medical Automatic Speech Recognition . . . 6
  - 1.3.1 German Medical Terminology . . . 7
  - 1.3.2 Austrian Standard German Pronounciation . . . 8
- 1.4 Research Questions . . . 10
- 1.5 Objectives and Contributions . . . 10

2 Background **13**
- 2.1 Transfer Learning . . . 13
- 2.2 Domain Adaptation . . . 15
  - 2.2.1 Domain adaptation in the context of Natural Language Processing and Automatic Speech Recognition . . . 15
  - 2.2.2 Domain Adaptation in the Context of Automatic Speech Recognition . . . 17
- 2.3 Types of Domain Adaptation in Natural Language Processing . . . 18
  - 2.3.1 Model- and Data-Centric Domain Adaptation . . . 19
  - 2.3.2 Availability of Classes in Target Domain . . . 20
- 2.4 Overview of Architectures used for DA . . . 21
  - 2.4.1 BERT . . . 21
  - 2.4.2 wav2vec 2.0 . . . 22
  - 2.4.3 Whisper . . . 23
- 2.5 Finetuning Methods . . . 25
  - 2.5.1 Hyperparameters . . . 25
  - 2.5.2 Challenges . . . 26
  - 2.5.3 PEFT . . . 27
  - 2.5.4 AdaLora . . . 27

3 Related Work **29**
3.1 Domain Adaptation in the context of Senior Citizen German 29
3.2 Domain Adaptation in the context of Child Speech 31
3.3 German Automatic Speech Recognition Datasets 33

4 Datasets **35**
4.1 Data Collection 35
4.2 Dictation Setting 35
4.3 Data Preparation 36
4.4 Statistical Analysis of Training Data 36
4.5 Dataset Structure 39
4.5.1 Separate Dataset for Testing on Standard German 39
4.6 Limitations 40
4.7 Ethical Considerations 40
4.8 Possible Applications 41

5 Baseline **43**
5.1 OpenAI's Whisper Model 43
5.1.1 Selection Considerations 46
5.2 Model Versions 46

6 Experiments **49**
6.1 Benchmark of Original Whisper Models 49
6.2 Domain Adpatation: Finetuning Experiments 51
6.2.1 Finetuning Whisper-medium 51
6.2.2 Impact of Training Data Volume on Model Performance . . 52
6.2.3 Finetuning Whisper-large-v3 53
6.3 Error Annotation 54

7 Results **57**
7.1 Benchmarking Original Whisper Model 57
7.2 Finetuning Results 58
7.2.1 Finetuning Whisper-medium 59
7.2.2 Impact of Training Data Volume on Model Performance . . 63
7.2.3 Finetuning Whisper-large-v3 66
7.3 Error Analysis 72

8 Discussion **81**

9 Limitations **83**

10 Conclusions **85**
10.1 Future Work 85

Bibliography **89**

List of Tables

1 Example of medical data transcriptions. 40

2 Training Settings of Finetuning Whisper-medium 52
3 Training Settings of Finetuning Whisper-medium with different Amounts of Training Hours 53
4 AdaLora Training Configuration 54

5 WER Results (%) ↓ of original Whisper models (with normalization) 57
6 WER Results (%) ↓ of original Whisper models (without normalization) 58
7 WER Results (%) ↓ for fine-tuned Whisper models 71

List of Figures

1 Difference between Traditional ML Processes and Domain Adaptation. Adapted from (Pan and Yang, 2010). 16

2 Distribution of Sentence Length of Training Data 37
3 Most Common Words in Training Data 38
4 Most Common Bigrams in Training Data 38
5 Distribution of Audio Lengths in Training Data 39

6 Overview of the Whisper Model. Figure by Radford et al. (2023) . . 44

7 Word Error Rate Explanation 50

8 Whisper-medium Finetuning: Evaluation Steps 59
9 Whisper-medium Finetuning: Training Loss 60
10 Whisper-medium Finetuning: Validation Loss 61
11 WER (%) ↓ improvements for Finetuned Whisper-medium (without normalization) on test set of MedASR-DE-AT 61
12 WER (%) ↓ improvements for Finetuned Whisper-medium (without normalization) on MedASR-DE-test 62
13 WER (%) ↓ improvements for Finetuned Whisper-medium (with normalization) on test set of MedASR-DE-AT 62
14 WER (%) ↓ improvements for Finetuned Whisper-medium (with normalization) on MedASR-DE-test 63
15 Impact on Training Data Volume on Performance of Whisper-medium on MedASR-DE-AT 64
16 Impact on Training Data Volume on Performance of Whisper-medium on MedASR-DE-test 65
17 Whisper-large-v3 Finetuning: Evaluation Steps 67
18 Whisper-large-v3 Finetuned: Training Loss 68
19 Whisper-large-v3 Finetuned: Validation Loss 68
20 WER (%) ↓ improvements for Finetuned Whisper-large-v3 (with normalization) on test set of MedASR-DE-AT 69
21 WER (%) ↓ improvements for Finetuned Whisper-large-v3 (with normalization) on MedASR-DE-test 70

22 WER (%) ↓ improvements for Finetuned Whisper-large-v3 (without normalization) on test set of MedASR-DE-AT 70
23 WER (%) ↓ improvements for Finetuned Whisper-large-v3 (without normalization) on MedASR-DE-test 71
24 Most Common Labels: Original Whisper-medium (MedASR-DE-AT) 73
25 Most Common MEDTERM Errors: Original Whisper-medium (MedASR-DE-AT) 74
26 Most Common Labels: Original Whisper-medium (MedASR-DE-AT) 75
27 Most Common Labels: Original Whisper-medium (MedASR-DE-test 76
28 Most Common MEDTERM Errors: Original Whisper-medium (MedASR-DE-test 77
29 Most Common Labels: Finetuned Whisper-medium (MedASR-DE-test) 78
30 Most Common MEDTERM Errors: Finetuned Whisper-medium (MedASR-DE-test) 78

1 Introduction

Automatic Speech Recognition (ASR) is the automatic transcription of vocal input Lu et al. (2020). ASR technologies convert spoken language into written text. Speech represents the most intuitive, effective, and favored means of human communication. We experience a greater comfort with using voice as an input method for machines compared to more rudimentary tools like keypads and keyboards. ASR systems enable this preference of communication by enabling voice based human-machine interaction (Malik et al., 2021). Furthermore, ASR systems can be of great help to people with disabilities who are limited in their ability to use conventional human-computer interaction methods (Semary et al., 2024). The objective of an optimal ASR system is to convert voice input into machine-readable data, enabling its use for subsequent machine-related tasks (Vadwala et al., 2017). For example, in medical ASR applications, a healthcare professional may dictate a patient's file, which is then automatically integrated into electronic health records EHR for future use (Avendano et al., 2022).

Over time, the demands and application contexts for ASR have evolved from simple to more challenging environments. There has been a transition from limited to extensive vocabularies, from speaker-dependent to speaker-independent systems, from read to spontaneous speech, and from quiet offices to noisy settings. Consequently, ASR techniques have significantly advanced over the past decades. Initially, basic template matching techniques like dynamic time warping were used during the 1950s to 70s. By the 1980s and 90s, the adoption of hidden Markov model-based statistical learning methods became prevalent, and more recently, deep learning techniques have become dominant (Lu et al., 2020). Thanks to the integration of these neural models and computational architectures that closely resemble human auditory and cognitive processing mechanisms, ASR systems have seen significant advancements. Modern ASR systems commence with the collection of speech through microphones, where the raw audio signal undergoes a conversion from analog to digital data to facilitate further processing. Once the data is digitized, these audio signals are pre-processed using techniques inspired by the auditory processing in humans. First, the speech signals get broken into various frequency components, often through Fourier Transforms (Oruh and Viriri, 2020). The Fourier Transform is a mathematical method that converts a time-domain function into its frequency-domain counterpart. This transformation reveals the various frequencies composing the original function, facilitating an in-depth

analysis of its frequency components. Essentially, it breaks down a waveform into its constituent frequencies, which could be compared to decomposing a musical chord into individual notes (Guan et al., 2021). Subsequently, these components get simplified into cepstral coefficients which offer a representation that is more resilient to variations in speech signals. The utilization of the Bark scale or Mel-frequency cepstral coefficients (MFCCs) is common in this context (Ayvaz et al., 2022). Cepstral coefficients, specifically (MFCCs), are a feature representation of sound based on the perceptual properties of human hearing. They are derived from the power spectrum of a signal, using a logarithmic transformation and a cosine transform, mapping these on a Mel-scale that approximates the human ear's response to different frequencies. The Bark scale, another psychoacoustic scale, similarly transforms frequency components to match human auditory perception but has slightly different frequency band allocations compared to the Mel scale. Both scales are used to ensure that the cepstral coefficients reflect how humans perceive sound, making them highly effective for speech and speaker recognition tasks in ASR systems (Nagarajan et al., 2020).

The core of ASR today involves feature extraction and acoustic modeling. Here, deep neural networks (DNN), particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), play a crucial role. These networks learn to map the complex relationships between the acoustic features and the linguistic units, like phonemes or words, of the speech. Recent advancements have seen the use of self-supervised learning models like wav2vec (Baevski et al., 2020), which further refine the feature extraction process by learning robust speech representations directly from unlabeled audio data (Shi et al., 2022).

Finally, the decoding process involves converting the recognized phonetic or feature sequences into legible text. This process uses language models, which are statistical models that predict the likelihood of sequence of words, to form coherent sentences. These models can be particularly complex, often leveraging the power of Transformer models, like OpenAI's Whisper (Radford et al., 2023) or other advanced architectures that manage contextual information over longer text sequences for more accurate prediction and generation of text from speech (Deng et al., 2022).

After more than fifty years of development, ASR technology has only recently begun to yield significant practical outcomes. These technical details highlight the sophisticated nature of contemporary ASR systems, which are increasingly capable of handling complex and natural speech in real-time applications. Notably, advancements in reducing the Word Error Rate (WER), with ASR systems getting close to human-like accuracy, have expanded ASR's practical uses (Wang et al., 2022).

1.1 Challenges in Automatic Speech Recognition

While ASR technologies exhibit improving accuracy over time, concerns about error rates persist. Variability in error incidence across different medical specialties and documentation types surfaced as an important consideration, emphasizing the need for tailored approaches in optimizing ASR precision (Lai, 2022). The performance of ASR systems can vary greatly depending on the language and the availability of relevant training data. Currently, training resources are available for only a limited number of languages, despite there being around 6,500 languages worldwide. For languages like German, especially in specialized domains such as medicine, the scarcity of datasets with both audio and textual data poses a considerable challenge (Graham and Roll, 2024).

Communication between humans and machines via speech faces significant challenges due to the vast variations in human speech patterns and accents. Inaccuracies can be introduced due to aspects such as differences in pronunciation speed, divergent speech patterns among speakers, and the emotional state that a user is in. These circumstances have the potential to significantly affect the performance of speech recognition systems (Semary et al., 2024). Additional complexities arise with factors like gender, dialect, speaking style, and speed (Malik et al., 2021). The diversity of dialects within natural language processing NLP presents difficulties due to the varied linguistic backgrounds of speakers, complicating common language detection (Hirayama et al., 2015). Moreover, spontaneous speech poses challenges as unpredictable content and pronunciation variations, potentially due to conditions like stuttering or limited speech abilities in children, directly impact ASR system performance (Kumar et al., 2020).

Another principal challenge for ASR is achieving reliable performance in noisy environments, where system efficacy is significantly diminished (Agrawal and Ganapathy, 2019). Despite numerous proposed algorithms for ASR, they often fail under real-world noisy conditions (Lee et al., 2016). In such environments, acoustical features are critically compromised, particularly in short-frame detection (Ming and Crookes, 2017). This refers to the process where the speech signal is divided into small segments or frames for analysis. In ideal conditions, these frames allow the ASR system to detect and analyze the acoustic properties of speech—like frequency and amplitude, assuming they remain relatively stable within each short interval. However, short-frame detection often struggles with accurately identifying or isolating acoustic features due to the reduced temporal resolution. This limitation compromises the ability to capture sufficient details within each frame, leading to potential inaccuracies in ASR (Jang et al., 2021). Furthermore, the presence of noise disrupts the effective representation of input signals in spectro-temporal terms, thereby obscuring the time-frequency correlations essential for analyzing speech signals. This interference directly impacts the initial

stages of speech signal processing where signals are decomposed into their frequency components via Fourier Transforms. Noise can mask or distort these frequency components, leading to less accurate conversions and representations (Ganapathy, 2017). While various noise reduction strategies exist, their effectiveness is limited unless the specific noises are previously identified. Noise signals, possessing diverse characteristics in real-world scenarios, pose a substantial challenge in creating a responsive database for DNN-based systems (Lee et al., 2018).

Environmental noise also results in the cocktail party problem, where overlapping speech from multiple speakers hinders accurate speech detection, a persistent issue in ASR (Chen et al., 2018). Furthermore, ASR performance is directly influenced by the quality of the capturing microphone. Low-quality microphones diminish ASR system performance (Gosztolya and Grósz, 2016), and background noise can cause distortions in channels. Additionally, the implementation of speech segregation requires more microphones than the number of sound sources (Dávila-Chacón et al., 2019). Therefore, there is a necessity for high-quality microphones for adequate performance (Agrawal and Ganapathy, 2019).

1.2 Medical Automatic Speech Recognition

Efficient and accurate documentation is imperative in modern healthcare, influencing clinical decision-making, patient care, and overall healthcare quality (Wurster et al., 2023). The integration of cutting-edge technologies into the field of healthcare continues to reshape traditional practices, offering innovative solutions to optimize workflows and enhance patient care. Among these technological advancements, ASR technology has surfaced as a promising tool within the domain of medical documentation, revolutionizing the way healthcare professionals dictate, transcribe, and manage patient information (Blackley et al., 2019). Studies show that the integration of ASR technology within medical documentation systems is a crucial development, promising transformative enhancements in the documentation process, as well as improvements in transcription accuracy and user contentment (Zuchowski and Göller, 2022; Blackley et al., 2020).

1.2.1 Challenges in Medical Automatic Speech Recognition

Kumar (2024) conducted an in-depth analysis of ASR systems employed in healthcare settings. In their research, they identified various challenges specific to the medical context, such as:

- *Medical Terminology:* Medical professionals frequently use specialized and context-specific language with specialized terms, abbreviations, and jargon

that are unique to the field. This includes complex anatomical terminology, diverse pharmacological names, and various procedural terms that may not be common outside of medical contexts. This often leads to transcription errors in speech recognition systems, which could negatively impact patient care.

- *Accents and Dialects:* The effectiveness of ASR systems used in healthcare is compromised when encountering accents and dialects different from those on which they were trained, resulting in transcription inaccuracies.

- *Sensitive Data and Privacy Concerns:* Healthcare data is highly confidential, governed by strict regulations. The deployment of speech recognition technologies thus raises significant security and privacy concerns.

- *Variability in Speaking Styles:* Healthcare environments often feature varied speaking styles, including differences in speed and syntax, which can reduce the accuracy of speech recognition systems.

- *Limited Context Understanding:* Despite improvements in contextual understanding, speech recognition systems still struggle with the complex dialogues typical in medical settings that require a deep understanding of the context and patient history.

- *Integration with EHR Systems:* Seamless integration with EHR systems is crucial yet challenging, often hindered by compatibility and interoperability issues.

- *Data Bias:* The performance of these systems can be limited by the representativeness of the training data, which, if biased, can diminish effectiveness across diverse patient groups.

- *Handling Errors and Misinterpretations:* Inaccuracies in speech recognition can lead to errors that, if not promptly corrected, might adversely affect patient outcomes.

- *Real-Time Processing:* The need for real-time processing in medical emergencies poses a challenge, as achieving both low latency and high accuracy can be difficult.

- *Ambient Noise and Acoustics:* Hospital settings are often noisy, which can disrupt speech recognition accuracy. Noise-reduction microphones are a potential mitigation strategy.

- *Variable Quality:* Without medical transcriptionists, physicians are tasked with ensuring record accuracy, leading to potential oversights such as misgendering, which are often not detected.

Kumar (2024) conclude that transcription errors present significant challenges in both types of speech recognition systems: front-end systems, which transcribe speech in real time, and back-end systems, which record speech for later processing (Aggarwal and Dave, 2011). These errors can potentially lead to the oversight of critical details. In back-end systems, poor grammar necessitates rigorous double checking by editors, especially if the system fails to recognize a specific user, thereby increasing operational costs due to the need for additional personnel such as editors and transcriptionists. Furthermore, the risk of losing or compromising patient data represents a serious concern. Additionally, the adoption of speech recognition technologies can be hindered by factors such as the age of physicians or their speech impairments, which may affect the system's accuracy and functionality.

The integration of ASR technology into the medical sector holds the promise of transforming healthcare documentation and communication. However, the complexities involved in medical and German ASR, especially when combined, highlight the need for ongoing research and development to enhance the accuracy, reliability, and cultural competence of these systems. Ensuring these systems can adequately handle the nuances of medical discourse and the specifics of the German language will be crucial for their success and acceptance in clinical environments.

1.3 German Medical Automatic Speech Recognition

ASR systems such as OpenAI's Whisper (Radford et al., 2023), wav2vec 2.0 (Baevski et al., 2020), or BERT-based speech models (Devlin et al., 2019) have been extensively developed for various applications, but their adaptation to the medical field in German-speaking regions has been limited due to the lack of specialized datasets containing the highly complex terminology used in German-speaking healthcare settings, which restricts the efficacy of these technologies (Wirth and Peinl, 2023; Liang et al., 2023). Furthermore, the need for datasets that consider the linguistic nuances of the Austrian Standard German pronunciation specifically has rarely been addressed, thus limiting the scope of existing solutions (Wirth and Peinl, 2022).

1.3.1 German Medical Terminology

The development of German medical terminology shows a dynamic linguistic evolution shaped by historical, cultural, and technical changes. Dobrić (2013) investigated and connected these transformations. In the following section, we discuss her findings, highlighting their significance and implications for the German medical language.

Hippocratic texts dating back to the 5th and 4th centuries BC are recognized as the earliest written documents in Western medicine. They mark the inception of the Greek period in medical terminology. Contrary to expectations that Latin would replace Greek medical terms following Roman dominance, the lack of a comparable Roman medical tradition and the prevalence of Greek physicians throughout the Roman Empire led to the adoption of Greek medical practices. Consequently, numerous medical terms of Greek origin persist in contemporary medical vocabularies, including those used in German.

During the early first century AD, *De Medicina*, a comprehensive overview of medical knowledge built on Greek sources, was written by the Roman aristocrat Aulus Cornelius Celsus. Celsus faced significant challenges in translating Greek medical terms into Latin, as many did not have direct Latin counterparts. To address this, he described symptoms in Latin but retained the original Greek terms within the text. This practice contributed to a blend of Greek and Latin in medical terminology that is evident in the German medical language today. The flexibility of Greek in forming new compound words often makes it preferable to Latin. The Greek prefix *hyper-* is commonly used over the Latin *super-* due to its broader utility in creating hybrid words. An example of the use of Greek in the German medical language is: *Hypertonie* combining the prefix *hyper-* with the Greek-derived *Tonie*, from *Tonus*, thus forming a term that precisely conveys an*extreme increase* in tone or tension.

A crucial transformation in medical language happened during the Renaissance, a period marked by widespread cultural revitalization. Paracelsus, an influential figure whose name suggests at least equality if not superiority to his predecessor Celsus, was instrumental in innovating the medical field. He differentiated himself by identifying new diseases and pioneering a unique approach to medicine. Paracelsus developed an extensive German medical terminology, crafting over 3,000 terms through new methods such as semantic shifts, derivations, and compound formations. He introduced names for diseases based on their geographic origins, such as *Franzosen* (The French) for syphilis, or their specific contexts, like *Bergsucht* (Addiction to Mountains) for miner's disease, which reflects a methodological reliance on topographical naming. For instance, *Hauptwe* (Headache) and *Halswe* (Sore Throat) were named for their respective locations of origin, the head and throat. While modern German has adapted these to *Kopfweh* and *Halsweh*, Paracelsus's impact

on medical nomenclature was relatively modest, as the medical field continued to prefer Latin and Greek terms. Despite this, his legacy persists in international chemical terminology with terms like *Alkohol* (Alcohol) and *Zink* (Zinc), the latter derived from miner's slang. Ultimately, Paracelsus's efforts to establish German as a language of medicine were not widely adopted until the 19th century when national languages began to play a more significant role in medical discourse.

Today, the predominant medical journals are mostly published in English, resulting in medical professionals worldwide having to adopt English as their primary language for international communication. As discussed, medical terminology was primarily derived from Latin and Greek, but with English now serving as the primary medium for medical discourse, new terms often include English words, such as *Bypass* or *Screening*, as well as hybrid terms that combine Greek or Latin prefixes or suffixes with English roots, like *Biofeedback*. Medical practitioners are faced with the option of directly incorporating these English terms into their vocabulary, like Celsus's adoption of Greek terms, or translating them into their native languages, following Paracelsus's approach with German. This dual practice in term creation, adopting foreign terms versus translating them, leads to an increase of synonyms within the German medical language. Dobrić (2013) concludes that, in theory, it is recommended to maintain consistency in term usage, as adherence to precise terminology is a crucial aspect of scientific language. This principle is also integral to established terminological guidelines aimed at enhancing clarity and reducing confusion in medical communication. However, these terminological principles, despite being widely recognized, are not always adhered to in medical practice.

The distinctive characteristics of German medical language discussed above pose significant challenges for ASR systems. The complexity of German medical terminology, with its deep roots in Greek and Latin, includes hybrid terms that combine elements from these classical languages with modern German and English, which complicates the ASR's ability to accurately recognize and process terms, especially when they include compound words or specialized prefixes and suffixes that may not be common in everyday language usage (Wirth and Peinl, 2022).

1.3.2 Austrian Standard German Pronounciation

The pronunciation differences between German as spoken in Germany and Austrian Standard German (ASG) pose significant linguistic distinctions that affect ASR tasks (Linke et al., 2023). ASG exhibits unique phonetic and phonological characteristics that diverge from Standard German, especially in conversational contexts. For instance, the GRASS corpus studies Schuppler et al. (2014b) indicate specific pronunciation rules prevalent in Austrian German, such as the lenition of fortis plosives and the spirantization of lenis plosives, which do not have direct counter-

parts in Standard German. The lenition of fortis plosives involves the softening of *fortis* (strong) plosive sounds. Fortis plosives are typically articulated with a burst of air and significant vocal effort, including sounds like *p* and *t*. In Austrian German, these sounds may undergo lenition; for example, a *t* might be pronounced with less force, making it sound slightly closer to a *d*, or it might be produced with more breathiness, resembling a soft fricative sound. Spirantization refers to the transformation of *lenis* (soft) plosives into fricatives. Lenis plosives, such as *b* and *d*, are typically voiced and softer than fortis plosives. In Austrian German, these sounds can be transformed into corresponding fricatives. For instance, the *b* in *Abend* (Evening) might be pronounced more like a *v* sound, resulting in a pronunciation that sounds more like *Avend*. Both processes lead to a softening of the typical plosive sounds, making them sound less sharp and explosive and more continuous or frictional. These phonetic shifts contribute to a distinctive Austrian German sound, affecting the clarity and identity of spoken words. Such changes can be challenging for ASR technologies, as they differ significantly from the Standard German pronunciation typically taught in training ASR systems (Linke et al., 2023).

In another study, Schuppler et al. (2014a) revealed that Austrian German often employs vowel monophthongization, where German diphthongs are simplified into a single, longer vowel sound. For example, the word *Nein* (No), typically pronounced with a diphthong as *nah-yen* in Standard German, is often pronounced as a simplified *nah* in Austrian German. Furthermore, they revealed various specific consonant realizations, such as the vocalization or deletion of *r* and the realization of the word-final syllable *-ig* as *ik* instead of the more open vowel sound used in Standard German. In Standard German, words ending in *-r* are pronounced with a clear *r* sound. In contrast, Austrian German tends to either vocalize or omit this sound. For instance, *Lehrer* (Teacher) is pronounced as *lehr-er* in Standard German but may be articulated as *lehra* or *lehr* in Austrian German, where the final *r* is either softened into a vowel sound or dropped entirely. In addition to that, the pronunciation of the word-final syllable *-ig* as *-ik* marks another phonetic distinction. In Standard German, this syllable typically has an *ich* sound, as in *König* (King), pronounced *kö-nich*. However, in Austrian German, this ending is pronounced sharply as *kö-nik*, substituting the softer *ch* sound with a clearer *k* sound.

These variations influence how speech is articulated in real life scenarios. Such phonetic and phonological shifts can lead to systematic discrepancies in ASR tasks, where models trained on Standard German data might misinterpret or fail to recognize Austrian pronunciations. Consequently, ASR systems require customized adaptations or specific training on Austrian German speech data to handle these differences effectively, addressing both the pronounced and more subtle divergences

in pronunciation that characterize the Austrian variant of the German language (Linke et al., 2023).

1.4 Research Questions

In this thesis, we aim to answer the following research questions:

Research Question (1)

How much effect does the domain adaptation of transformer-based ASR models, such as OpenAI's Whisper, with German medical speech data have on the model's performance?

We seek to assess the impact of domain adaptation on the performance of transformer-based ASR models, specifically focusing on OpenAI's Whisper (Radford et al., 2023). To this end, we have developed a German medical dataset within the surgery domain. This dataset encompasses a range of surgical terminology and linguistic intricacies that are typical in medical settings. By training and finetuning Whisper models using this specialized dataset, we aim to determine whether such domain-specific adaptation can enhance the model's accuracy and effectiveness in medical ASR scenarios.

Research Question (2)

What is the impact of dialectal variations, such as Austrian Standard German versus Standard German pronunciation, on the performance of Whisper models?

We finetune the Whisper models using our custom medical dataset, which was dictated by two speakers of Austrian Standard German. Additionally, we conduct tests on Standard German pronunciation to evaluate any potential performance discrepancies or variations between the dialects.

1.5 Objectives and Contributions

The evolution of ASR technologies has brought significant advancements in various fields, including healthcare. However, as discussed in Section 1.3, despite these advancements, the application of ASR in medical settings, particularly within the German-speaking regions where Austrian Standard German pronunciation is used, faces distinct challenges. In this thesis, we aim to address these challenges by applying domain adaptation and finetuning approaches on Whisper models

(Radford et al., 2023) to enhance the accuracy and reliability of ASR systems in medical contexts. The overall goal is to develop an optimized ASR model that can effectively manage the nuances of German medical terminology pronounced in Austrian German.

In order to achieve this goal and answer the research questions, we will conduct a series of different experiments. Our main contributions are outlined below:

Benchmark of Original Whisper Models We benchmark the original Whisper models by calculating the WER on our dataset's test split. This evaluation serves as a basis for comparison, to measure the effectiveness and improvements achieved through our subsequent finetuning experiments. We present the WER results of the original Whisper models in Section 6.1.

Development of a Surgery-Specific German Medical Dataset One of the primary contributions of this thesis is the development of a specialized German medical dataset tailored to the surgical domain. This dataset is unique in its incorporation of Austrian Standard German pronunciation, including linguistic nuances specific to Austria. The creation of such a dataset is vital as it fills a critical gap in existing ASR training resources, which, as of now, lack specialized medical content. The creation of our dataset is presented in Chapter 4.

Finetuning of Whisper-Medium and Whisper-Large-V3 Models A significant portion of this research involved the finetuning of two configurations of the Whisper model: Whisper-medium and Whisper-large-v3. We used the newly developed medical dataset to finetune both models with the goal to refine their transcription capabilities in a medical context. We present our finetuning experiments in Section 6.2.1.

Evaluating the Impact of Training Data Volume on Model Performance Additionally, we conduct experiments to assess how the amount of training data influences the performance of the Whisper-medium model. We finetune the model with increasing volumes of data: from 1 hour to 9 hours, which constitutes the full training set, in one-hour increments. We evaluate the performance on the validation split of our dataset. These results are shown in Section 6.2.2.

Evaluation and Comparison of Model Performance After the finetuning experiments, we evaluate the adapted models to assess their performance improvements. We measure the models' performance using WER evaluation and compare them against the original Whisper models. This is illustrated in Section

6.2. The results offer important insights into the benefits and limitations of domain adaptation in ASR technology and are shown in 7.2.

Error Annotation and Analysis Additionally, we conduct an extensive error annotation and analysis using Label Studio [1]. We perform a systematic review of the transcriptions generated by both the original and finetuned models, identifying and categorizing errors, which is presented in Section 6.3. Such a detailed error analysis helps to identify specific weaknesses of the transformer models and provides a direction for further model refinement.

Documentation of Findings and Recommendations for Future Work Finally, this thesis documents all findings in a detailed manner, providing a clear depiction of the research outcomes and challenges encountered during the model adaptation process. Furthermore, we offer recommendations for future research in this area in Section 10.1.

This thesis contributes to the ongoing advancement of ASR technology, particularly in the field of domain adaptation for specific contexts like the German medical domain, with an emphasis on Austrian Standard German pronunciation. Through these contributions, we not only promote advancements in the field of ASR in terms of technical development and application but also establishes a foundation for future research that could lead to more robust and accurate ASR systems in healthcare and other critical sectors.

[1] https://labelstud.io/

2 Background

In the following chapter, we outline the fundamental concepts and methodologies integral to transfer learning, emphasizing domain adaptation, model architectures, and finetuning methods. We aim to establish a solid theoretical foundation and discuss practical applications, which are crucial for the subsequent sections of this thesis.

2.1 Transfer Learning

Machine learning techniques predict future outcomes by employing machine learning models trained on gathered data that may either be labeled or unlabeled. It is essential to differentiate between the two primary types employed in transfer learning: statistical and neural models. Statistical models are based on statistical techniques such as linear regression, logistic regression, or decision trees. They demonstrate proficiency through their ability to interpret smaller datasets (Shah et al., 2020). In contrast, neural models, or deep learning models, utilize layered neural networks capable of learning complex patterns and representations from large amounts of data. These models excel in tasks requiring the extraction of intricate features from raw data, such as image and speech recognition (Mehrish et al., 2023). Each model offers unique advantages and challenges for transfer learning, affecting their suitability based on the source and target tasks and data availability. Nevertheless, these models often encounter challenges during the training process due to a lack of labeled data, as well as issues related to imperfect datasets or inaccurate labels (Zhu, 2008). One possible solution for solving these problems is transfer learning (TL) as it enables training across differing source and target domains, tasks, and data distributions (Pan and Yang, 2010).

Transferring knowledge across different domains represents a more complex challenge. Achieving cross-domain transfer is a key objective in transfer learning, as it opens up the possibility of applying knowledge between tasks that are considerably dissimilar (Taylor and Stone, 2007). This approach is inspired by human learning, where knowledge gained from previous experiences is applied to new, unrelated tasks more efficiently or effectively. To give an example, being able to speak Italian might facilitate learning another Romance language, such as Spanish, or, learning to recognize a palm tree may be helpful when trying to identify a Yucca

plant. The concept was initially discussed in 1995 at a NIPS workshop called "Learning to Learn", where the focal point was the necessity of lasting machine learning approaches that reuse existing knowledge and improve over time, essentially learning how to learn (Pan and Yang, 2010). The acronym NIPS refers to the Conference on Neural Information Processing Systems, a prestigious conference within the fields of computational neuroscience and machine learning. In response to evolving disciplinary boundaries and to reflect a broader scope beyond neural processing systems, NIPS was rebranded to NeurIPS, the name by which the conference is recognized today (Neu). Since then, transfer learning has gained attention with various terms such as life-long learning, knowledge transfer, and knowledge consolidation, among others (Thrun and Pratt, 1998). Multitask learning, a related concept, focuses on simultaneously learning multiple tasks to identify shared features that benefit each task. However, transfer learning is distinct in its focus on optimizing performance on the target task by applying knowledge from source tasks, without the necessity of learning all tasks at once (Pan and Yang, 2010).

In 2005, the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency (DARPA)'s Information Processing Technology Office (IPTO) (Gasser et al., 2005) outlined a specific objective for Transfer Learning: to enhance systems' capabilities to apply learned skills and knowledge from previous tasks to new, different tasks. This marked a shift towards emphasizing the application of acquired knowledge to a target task, putting the priority on the asymmetrical relationship between source and target tasks in transfer learning (Shi et al., 2012). Contrary to traditional approaches, where the aim is to learn each task independently, transfer learning methods focus on applying knowledge acquired from previous tasks to a new task, which proves particularly beneficial when the new task has limited high-quality data available for training (Pan and Yang, 2010).

Pan and Yang (2010) categorized Transfer Learning into three distinct forms: inductive, transductive, and unsupervised TL, each characterized by the relationship between the source and target tasks and domains, as well as the nature of the data used. Inductive TL involves training a model on a source task and then adapting it to a different target task. This adaptation process does not necessarily require the training data to be labeled, but it is crucial for the target data to be labeled. Transductive TL, on the other hand, operates based on the assumption that the source and target tasks are identical but are applied across different domains. Here, the critical factor is that while the source data must be labeled, the target data does not contain labels. Finally, unsupervised TL is characterized by a lack of labeled data in both the source and target domains, with the tasks across these domains also differing. This form of TL presents a significant challenge, as it requires the

model to learn and transfer knowledge without the guidance of explicit labels in either the learning or application phases (Niu et al., 2020).

In the context of transfer learning, a critical strategy for addressing the challenge of applying knowledge across different domains, particularly when the source and target data distributions diverge significantly, is domain adaptation (Farahani et al., 2021). Domain adaptation focuses on modifying the transfer learning approaches to bridge the gap between diverse domains. This is particularly relevant when dealing with scenarios where the source and target tasks are similar but are executed in different environments. By leveraging the principles outlined in transfer learning, domain adaptation techniques aim to minimize domain discrepancies and to enhance the model's ability to generalize from the source domain to the target domain effectively (Singhal et al., 2023). In the following section, we will explore how domain adaptation (DA) approaches are applied to ensure the effectiveness of knowledge transfer, despite the challenges posed by differing domain characteristics.

2.2 Domain Adaptation

As discussed, domain adaptation represents a specific form of transfer learning, particularly within the framework of transductive learning. Similarly to transductive learning, in domain adaptation source and target task remain consistent, while the domains may differ. However, whereas transductive learning assumes that the target data is unlabeled, domain adaptation does not necessarily make this assumption (Niu et al., 2020). Hence, domain adaptation is a technique in the field of machine learning that aims to adapt a model trained on a source domain to perform well on a different target domain, addressing the challenge of differing data distributions in training and testing environments. This approach is critical for enhancing the generalizability of machine learning models across varying contexts, such as Computer Vision (CV) or NLP (Singhal et al., 2023).

Figure 1 illustrates the distinctions in the learning approaches between traditional machine learning methods and DA strategies. As shown here, traditional machine learning tactics involve training models separately for distinct domains. Conversely, DA methods enable the application of a pre-trained model's knowledge to new domains, previously unseen by the model.

2.2.1 Domain adaptation in the context of Natural Language Processing and Automatic Speech Recognition

In this section, we first present formal definitions of the terms *domain* and *task* as they pertain in the field of domain adaptation. These definitions are based on the

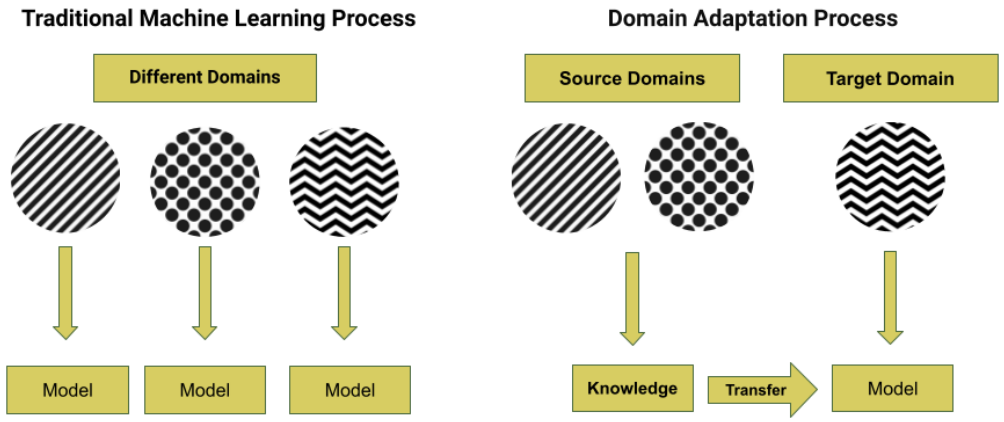

Figure 1: Difference between Traditional ML Processes and Domain Adaptation. Adapted from (Pan and Yang, 2010).

work of Bashath et al. (2022) and serve to establish a foundational understanding for further discussion on domain adaptation in the field of NLP, and, more precisely, within the context of ASR.

Domain A domain D is defined as a tuple $D = \{x, P(X)\}$, where x represents the set of all instances, X is an instance, and $P(X)$ denotes the marginal probability distribution, which refers to the probability distribution of one or more variables within a set, derived by summing or integrating over the probabilities of the other variables in the set (Deisenroth et al., 2020), over these instances. Each instance within this set, denoted as $X \in x$, exemplifies an individual case within the domain (Bashath et al., 2022).

Task Given a domain $D = \{x, P(X)\}$, a task T is characterized by the tuple $T = \{y, f\}$, with y being the label space and f the prediction function. This function, $f : x \rightarrow y$, maps instances to labels and is encapsulated by the conditional probability distribution $P(Y|X)$. Consequently, a task can be represented as $T = \{y, P(Y|X)\}$, signifying the association between instances and their corresponding labels (Bashath et al., 2022).

However, it should be highlighted that, within the NLP community, the term

domain is often defined in a somewhat flexible manner. It typically refers to a specific collection of linguistic data, characterized by commonalities such as topic, style, genre, or linguistic register. For example, the news domain is characterized by a formal style and topics related to current events, whereas the IT domain is marked by technical jargon and topics related to technology and software. While the legal domain is distinguished by its use of legal terminology, complex sentence structures, and topics related to laws and regulations, the medical domain features specialized vocabulary and clinical information. It includes a wide range of terms for different specialties, such as oncology, cardiology or surgery, and there is an extensive use of abbreviations and acronyms. Additionally, it comprehends a mixture of technical and non-professional terms, that often derive from foreign languages such as Latin or Greek (Djalilova et al., 2023).

Furthermore, in NLP, the terms source domain and target domain are commonly used to describe, respectively, the specific dataset on which a machine learning model is trained and a different dataset with a distinct distribution used for testing the model (Ramponi and Plank, 2020).

Domain adaptation is widely used in a vast variety of research areas, such as Computer Vision, Healthcare and Biomedical Engineering or Robotics (Kouw and Loog, 2018). However, the challenge of transferring knowledge from one domain to another is especially relevant in NLP, where it is common to have abundant labeled data in the source domain, like, for example, newspaper articles, but require a model that excels in a different target domain, say, the healthcare domain (Daumé, 2007). Domain adaptation provides a vast assortment of techniques that aim to enhance the performance of NLP models in situations where the target domain has scarce or no training data. These techniques facilitate bridging the gap between source and target domains, thereby contributing to the development of increasingly effective NLP applications (Kouw and Loog, 2018)

2.2.2 Domain Adaptation in the Context of Automatic Speech Recognition

In the context of ASR, the domain D can be exemplified as a tuple $D = \{x, P(X)\}$, where x represents the set of all possible audio recordings. Each instance in this set, denoted as $X \in x$, corresponds to an individual audio recording. $P(X)$ denotes the marginal probability distribution over these instances, reflecting the likelihood of occurrence of each audio recording in the real world. This distribution is influenced by factors such as word prevalence, terminology, accents, and background noise levels (Bashath et al., 2022).

Given the domain D, the task T for ASR can be characterized by the tuple $T = \{y, f\}$, where y is the label space representing all possible transcriptions of the

audio recordings into text. Each label in this set, denoted as $Y \in y$, corresponds to a specific transcription of an audio recording. The prediction function $f : x \rightarrow y$ maps instances (audio recordings) to labels (text transcriptions), encapsulating the process of automatically recognizing spoken words and converting them into corresponding text. Therefore, the task T can also be represented as $T = \{y, P(Y|X)\}$, signifying the association between instances (audio recordings) and their corresponding labels (text transcriptions) through the conditional probability distribution $P(Y|X)$. This distribution models the likelihood of a transcription given an audio recording considering factors such as pronunciation variations, speech speed, and background noises that affect transcription accuracy.

Based on these definitions, within the field of ASR, the domain involves understanding the characteristics of spoken language captured in audio recordings, including variations in accents, terminology, and environmental noise. The task, on the other hand, involves accurately predicting the textual transcription of spoken words in an audio recording, a process requiring recognition and interpretation of speech signals within the spoken language context.

2.3 Types of Domain Adaptation in Natural Language Processing

Domain adaptation in NLP can be classified into five primary learning paradigms that align with the typical three-part classification of machine learning models: supervised, unsupervised, semi-supervised, zero-shot, and few-shot learning. All these paradigms can be distinguished by the availability and utilization of labeled data in the target domain.

Supervised domain adaptation assumes the availability of a substantial amount of labeled data in the target domain. The primary strategy involves retraining or finetuning a pre-trained model, which had originally been trained on the source domain, using the labeled data from the target domain. This approach is effective in scenarios where obtaining labeled data in the target domain is feasible, despite the additional costs or efforts required. The key advantage of supervised domain adaptation is its potential to achieve high performance on the target task, given the direct guidance from the target domain labels (Plank, 2011).

Unsupervised domain adaptation is applied in contexts where no labeled data is available in the target domain. The focus here is on leveraging the large amount of unlabeled data in the target domain alongside the labeled data in the source domain. Techniques in this category often involve aligning the feature distributions of the source and target domains, making the assumption that a model performing well on the source domain will also perform well on the target domain if the

domain discrepancy is minimized. Strategies include self-training, mostly used in statistical machine learning (Amini et al., 2022), as well as domain-invariant feature extraction and adversarial training, often employed in neural machine learning settings (Gallego et al., 2021), among others. Unsupervised domain adaptation is particularly valuable in situations where labeling data in the target domain is impractical or too expensive (Plank, 2011).

Semi-supervised domain adaptation falls between supervised and unsupervised paradigms. It assumes that a small amount of labeled data and a large amount of unlabeled data are available in the target domain. This approach aims to exploit the strengths of both supervised and unsupervised methods by using the limited labeled data to guide the learning process while also leveraging the abundant unlabeled data to facilitate domain adaptation. The aim is to incrementally improve the model's performance on the target domain by iteratively refining its understanding of the target domain data (Plank, 2011).

Zero-shot transfer learning (Radford et al., 2021; Zhai et al., 2022) enables a model to tackle tasks it has not explicitly encountered during training. It differs from zero-shot learning in that it incorporates some supervised elements during the pre-training phase. However, zero-shot transfer learning avoids using supervised examples in the transfer phase. This capability stems from the model's ability to utilize its pre-trained knowledge to infer characteristics of new categories, making it especially useful in contexts where collecting extensive labeled data is not feasible (Pham et al., 2023).

Few-shot transfer learning is a machine learning approach that bridges the gap between traditional supervised learning and zero-shot learning by enabling models to adapt to new tasks with a minimal amount of labeled data. Typically, after pre-training on large, diverse datasets, these models are finetuned with only a few labeled examples per class in the target domain. This method utilizes the knowledge gained during the pre-training phase to effectively recognize patterns or make predictions in new contexts with limited additional training. Few-shot transfer learning is particularly useful in scenarios where acquiring extensive labeled data is impractical, thus providing a flexible and efficient way to extend the applicability of machine learning models to a broader range of tasks (Guo et al., 2020).

2.3.1 Model- and Data-Centric Domain Adaptation

Apart from classifying domain adaptation approaches based on the availability of labeled target data, Chu and Wang (2018) divide domain adaptation methods, specifically in the field of NLP, into two categories: model-centric and data-centric. The data-centric strategy lays the focus on the particular data being used for training and pre-training methods rather than customizing models for domain adaptation. This strategy employs a range of data sources, including domain-

specific monolingual datasets, artificially generated datasets designed to mimic domain-specific conditions, and parallel datasets that offer comparable materials in multiple languages. Conversely, the model-centric strategy centers on enhancing machine learning models for improved performance in specific domains. For neural machine learning models, this involves finetuning the models through changes to their training parameters to better match domain requirements, adjusting their structure to increase their efficacy in domain-specific contexts, or tweaking the algorithms used for decoding to yield more accurate domain-specific results (Chu and Wang, 2018).

Ramponi and Plank (2020) added a third differentiation to domain adaptation methods in NLP, the hybrid approach. This hybrid method combines elements of both data-centric and model-centric strategies. It aims to leverage the strengths of both strategies to achieve more effective domain adaptation, optimizing both the data provided to the model and how the model learns from this data.

2.3.2 Availability of Classes in Target Domain

Farahani et al. (2021) divide domain adaptation into four different groups: open set, partial, universal, and closed set domain adaptation. Open set domain adaptation allows for the existence of unique classes in the target domain not present in the source, requiring the model to recognize and handle new, unseen categories (Busto and Gall, 2017). Partial domain adaptation functions under the premise that the target domain contains a subset of the classes found in the source domain, with the challenge being to focus on the relevant shared classes (Cao et al., 2018). Universal domain adaptation aims to generalize across domains without prior knowledge of label sets, dealing with shared, unique, and outlier classes across both source and target domains (You et al., 2019). In contrast, closed set domain adaptation, the most common kind, assumes that the source and target domains share the exact same classes, focusing on bridging the domain gap while maintaining identical label sets (You et al., 2019).

In this thesis, we employ a supervised, closed set domain adaptation approach to OpenAI's pre-trained transformer model Whisper. The methodology is hybrid, combining both data-centric and model-centric strategies to enhance the model's performance within the medical domain. The incorporation of German medical domain data will help the model to accurately process German medical terminology, while adjustments to the learning parameters optimize its performance for the specialized medical context. This approach ensures a comprehensive domain adaptation, maintaining class label integrity across domains. As discussed, domain adaptation emphasizes the reuse of knowledge gained from prior tasks for new ones, which is especially beneficial when there is not enough high-quality training

data for the new task. This concept holds significant importance for this thesis for two main reasons. First, ASR training data in the German medical domain is limited (Wirth and Peinl, 2022). Second, Whisper is trained on multilingual and multitask data, which enables the use of extensive pre-existing knowledge to enhance performance in this specialized context (Radford et al., 2023).

2.4 Overview of Architectures used for DA

In the following section, we will discuss the most prevalent model architectures and frameworks for domain adaptation that have proven effective for tasks like NLP and ASR. The decision which models to include is drawn from analyzing a wide array of research articles and reviews that collectively showcase the utility and adaptability of these models in overcoming the challenges posed in domain adaptation.

2.4.1 BERT

Prominent for domain adaptation in NLP are pre-trained transformers, such as BERT and its variants. BERT is a multi-layer bidirectional transformer encoder that was introduced by Devlin et al. (2019). BERT can be adapted for various domain adaptation types, but it is particularly effective in closed set domain adaptation. However, its robust pre-trained model also provides a strong basis for adapting to new, even unseen classes in the target domain. BERT is pre-trained using a masked language modeling task where 15% of the input tokens are randomly masked, replaced variably with a special [MASK] token, a random token, or left unchanged, and the model learns to predict the original tokens from these altered inputs to develop a deep bidirectional representation. BERT is frequently used in unsupervised domain adaptation, since it can use its pre-trained knowledge to adapt to new domains without requiring labeled data from those domains. After pre-training, the model can be finetuned with just an additional output layer to create state-of-the-art models for a vast range of NLP tasks, such as Text Classification, Question Answering, Named Entity Recognition (NER), or ASR (Devlin et al., 2019).

As mentioned, BERT's effectiveness in domain adaptation and transfer learning comes from its deep understanding of language context, learned during pre-training. This allows it to adapt to a multitude of tasks and domains with minimal finetuning. Within the context of NLP, BERT has shown exceptional performance in tasks such as named entity recognition, sentiment analysis, and question answering, displaying its versatility and power as a model for improving NLP applications across various domains (Karouzos et al., 2021).

2.4.2 wav2vec 2.0

While BERT proves highly effective for NLP tasks involving written language, there are other models more tailored to tackle domain adaptation in the context of ASR, such as wav2vec 2.0 (Baevski et al., 2020). Wav2vec 2.0 was designed for unsupervised learning of speech representations using a convolutional neural network and a self-attention mechanism and was introduced by Facebook AI in 2020. By leveraging raw audio waveforms from multiple languages, wav2vec 2.0 learns a universal set of quantized latent speech representations shared across languages. This cross-lingual pre-training approach enables the model to capture nuanced speech features applicable across various linguistic contexts. Wav2vec 2.0 can be used for various ASR tasks, from traditional transcription to more complex scenarios involving multilingual and low-resource language speech recognition. Its ability to learn rich, language-independent representations makes it an important tool for developing robust, adaptable ASR systems capable of handling diverse and challenging acoustic environments (Baevski et al., 2020).

It can adapt to new speech domains or languages by finetuning on smaller, domain-specific datasets, thus addressing the problem of the availability of few labeled data in certain domains. This cross-lingual representation learning proves especially beneficial for low-resource languages, significantly reducing WERs by leveraging similarities with high-resource languages (Conneau et al., 2021).

Wav2vec's pre-training methodology makes it suitable for domain adaptation. The model's versatility allows it to be used for universal domain adaptation, especially due to its unsupervised learning capabilities which enable it to handle varied speech representations without predefined label sets. It can also be used in partial domain adaptation, given its application to target domains that include subsets of the source domain's classes. As mentioned above, wav2vec 2.0 employs an unsupervised learning approach during pre-training, however, for domain adaptation, it may also be applied in semi-supervised and self-supervised learning scenarios, utilizing both unlabeled data and a small amount of labeled data in the target domain (Hwang et al., 2022).

Domain adaptation approaches using wav2vec 2.0 are predominantly data-centric due to the model's reliance on raw audio data for learning. However, when adapting to specific domains, a hybrid strategy may be employed, adjusting the model to refine its performance on domain-specific tasks (Hsu et al., 2021).

XLSR-53 XLSR-53 (Conneau et al., 2021) extends the wav2vec 2.0 framework to multiple languages for pre-training. It was designed to learn universal speech representations by training on a vast dataset of unlabeled speech from various languages. Specifically, XLSR-53 was trained on 56,000 hours of speech data across 53 languages, inlcuding a diverse set of linguistic environments from the MLS

(Pratap et al., 2020), CommonVoice (Ardila et al., 2019), and BABEL (Punnakkal et al., 2021) datasets.

The architecture of XLSR-53 follows the wav2vec 2.0 design, featuring a convolutional feature encoder that maps raw audio into latent speech representations. These representations are then processed by a Transformer network, which outputs contextual representations. The model employs a contrastive loss function during training. This means to distinguish correct latent representations from a set of negative samples within masked regions of the input audio (Conneau et al., 2021).

XLSR-53 achieves significant performance improvements in ASR tasks across a variety of languages. By leveraging large-scale, unsupervised pre-training across languages, it demonstrates superior phoneme error rate reductions and WER improvements compared to monolingual baselines and prior state-of-the-art multilingual systems. The model's effectiveness is particularly notable in low-resource language scenarios, where it helps to mitigate the scarcity of labeled data by transferring learned representations from high-resource to low-resource languages (Conneau et al., 2021).

XLS-R XLS-R (Babu et al., 2022) is an advanced cross-lingual speech representation learning model, building upon the foundation set by its predecessor, XLSR-53, and extending the wav2vec 2.0 architecture. Unlike XLSR-53, which, as discussed, was trained on 56,000 hours of speech from 53 languages, XLS-R encompasses a more expansive training corpus with approximately 436,000 hours of unlabeled speech data across 128 languages. This expansion includes a broader array of both high and low-resource languages, significantly enlarging the training corpus beyond the datasets previously utilized. The enhanced scale of training allows XLS-R to achieve superior performance benchmarks on a variety of speech processing tasks, including speech recognition and speech translation across diverse languages and domains. The model's architecture effectively employs large-scale, self-supervised learning techniques to generate robust multilingual speech representations. Consequently, XLS-R excels in generalizing across different linguistic environments, particularly benefiting low-resource languages through effective cross-lingual transfer, thus improving performance where labeled data is scarce (Babu et al., 2022).

2.4.3 Whisper

OpenAI's Whisper model (Radford et al., 2023) emerged as a state-of-the-art model for ASR tasks, where text as well as spoken language need to be handled. Whisper is a pre-trained, end-to-end transformer, designed for ASR. Utilizing a large and diverse dataset sourced from the internet, it incorporates a simple

encoder-decoder architecture that enables it to recognize and transcribe speech from various domains, languages, and acoustic conditions (Radford et al., 2023). This versatility makes Whisper particularly suitable for domain adaptation and transfer learning in ASR. Its ability to be finetuned on domain-specific datasets allows it to adapt to new contexts, such as medical conversations or air traffic control communications, improving its performance in these specialized areas.

While Radford et al. (2023) do not provide a detailed breakdown of each specific source or the exact composition of the model's training data, they state that it was trained on a comprehensive dataset that included 680,000 hours of diverse audio tasks. This dataset encompassed English transcripts, translations from multiple languages into English, transcriptions in languages other than English, and exclusively instrumental audio, all of which were collected from online sources. This vast training makes Whisper a robust language model base that proves beneficial mainly for closed set domain adaptation. However, its capacity for further training makes it also adaptable to partial domain adaptation scenarios, where it may encounter a subset of known classes within a new domain. As Whisper benefits significantly from labeled data in the target domain for finetuning, the most commonly used approach is supervised domain adaptation (Radford et al., 2023). This is especially true when adapting to a highly specialized domain like German medical ASR, where accurate transcriptions are critical.

Recent studies exploring domain adaptation strategies employing the Whisper model, such as Pekarek-Rosin and Wermter (2023); Jain et al. (2023), predominantly utilize hybrid approaches, incorporating data-centric methods by training on domain-specific data and model-centric methods through finetuning model parameters to optimize performance in the target domain. This approach was inspired by advances in the NLP field and demonstrates significant WER reductions on unseen datasets, demonstrating Whisper's effectiveness in domain adaptation for a variety of ASR tasks (Liao et al., 2023).

Whisper Architecture Since we focus on adapting Whisper models to a new domain, we will now discuss the loss functions employed throughout the domain adaptation process. Whisper's training pipeline includes two stages: multitask pre-training and finetuning (Radford et al., 2023). In the context of domain adaptation using the Whisper model, loss functions play a critical role throughout the two-stage training process.

During pre-training, the model is exposed to various tasks such as translation and ASR across a multitude of different languages and uses the cross-entropy of the next word but conditioned on the audio as loss (hug, 2024). The primary objective here is to learn useful representations from a large corpus of speech data and to build a robust model foundation. However, the model's optimization during

pre-training is not directly tied to any specific downstream task. Instead, it aims to capture general acoustic and linguistic features (Radford et al., 2023).

Subsequent to this phase is the finetuning stage, where Whisper undergoes optimization for domain-specific data. Here, the focus shifts to minimizing losses that are customized to the target domain, which might involve the introduction of domain-specific modifications or the inclusion of special tokens to better guide the model in recognizing and accurately transcribing specialized terminology pertinent to the domain in question. These tokens can act as markers or signals to the model, emphasizing the importance of certain terms, phrases, or acoustic features unique to the domain (Ma et al., 2023).

These domain-specific modifications to the training process are essential for finetuning Whisper to achieve high accuracy and reliability in specialized contexts. By incorporating them, the model is better equipped to handle the unique challenges presented by the target domain, leading to improved transcription accuracy and a greater ability to generalize from the broad knowledge acquired during pre-training to the specific demands of the new domain.

2.5 Finetuning Methods

Finetuning is a key technique in machine learning that involves additional training and adaptation of pre-trained models to new datasets in order to tailor them to specific tasks or domains. This process is essential for harnessing the general capabilities of pre-trained models and customizing them to specialized domains (Zhang et al., 2022a). Finetuning is particularly beneficial for domain adaptation(Lee et al., 2023), such as adapting the Whisper model to the German medical context. It allows the model to learn domain-specific terminology and stylistic distinctions, which helps it to improve its performance in medical transcription tasks. As discussed in Section 2.1, this approach is more efficient than training a model from scratch, since it requires fewer data and computational resources to achieve high performance.

2.5.1 Hyperparameters

Several hyperparameters can be adjusted during the finetuning process to optimize a pre-trained model for specific tasks. These adjustments are essential for adapting the model to new domains and enhancing its performance on specialized datasets(Usmani et al., 2023).

The learning rate is one of these imperative parameters. This rate affects how much new data influences the existing information, effectively determining how quickly the model adapts to the new domain-specific data. In other words, it

represents the pace at which a machine learning model learns. During finetuning, the learning rate is typically set lower than in pre-training so the model can adjust its weights rather than learn from scratch, thereby avoiding overfitting on the smaller dataset (Jin et al., 2023).

The size of the training batches also influences the model's learning dynamics. Larger batch sizes tend to stabilize the training process and help the model converge more smoothly by providing a more accurate gradient estimate. However, they require more memory and computational resources. Conversely, smaller batches may lead to better generalization on unseen data due to their inherent noise, which can act as a form of regularization (Oyedotun et al., 2023).

Gradient accumulation is another technique used to handle limitations in computational resources. It accumulates gradients from multiple mini-batches before updating the model's weights. This method is crucial when training with limited GPU memory, since it allows for the use of larger training batch sizes than memory constraints would normally permit, enabling more stable and reliable weight updates (Huang et al., 2023).

An additional training parameter is the number of warmup steps. These are steps taken at the beginning of the training process, where the learning rate gradually increases from a low to a higher value. They are crucial for maintaining training stability. This method is particularly beneficial in preventing the model from prematurely converging to suboptimal solutions when starting with a higher learning rate (Gupta et al., 2023).

2.5.2 Challenges

Although finetuning offers considerable benefits, it also carries associated risks. One critical consideration is the balance between generalization and specialization. It is imperative to prevent overfitting the model to the nuances of the training data, which can impair its performance on broader contexts. This balance is important for ensuring that the model is not only accurate but also versatile in handling various domains, including medical scenarios (Ju et al., 2022). Employing strategies such as cross-validation, using a diverse set of training examples, and integrating various regularization techniques can mitigate this risk (Zhang et al., 2022b).

Additionally, finetuning large pre-trained models entails considerable computational expenses, since these models typically contain billions of parameters. To give an example, the Whisper large-v3 model [2] contains 1.55 billion parameters. The extensive scale and high computational demands of these large language models present significant challenges in adapting them for specific domains or tasks, especially on hardware platforms with limited computational capabilities (Shi et al.,

[2] https://huggingface.co/openai/whisper-large-v3

2023). To address this issue, Parameter-Efficient finetuning (PEFT) was developed (Houlsby et al., 2019a; Pfeiffer et al., 2020; Brown et al., 2020; Li and Liang, 2021).

2.5.3 PEFT

Parameter-Efficient finetuning (PEFT) is a useful strategy for adapting large pre-trained models with minimal computational cost. Specifically, PEFT involves the modification of parameters in pre-trained large models to suit particular tasks or domains, while minimizing the introduction of additional parameters and the computational resources required. In this process, only a small subset of parameters gets updated while keeping the majority frozen, thus saving time and resources during model training (Han et al., 2024).

Several PEFT techniques have demonstrated effective performance, while employing different strategies to adapt large pre-trained models with minimal computational overhead. Among these techniques are Adapters. They involve integrating small trainable modules between the layers of a pre-trained model. These modules are designed to selectively modify a limited number of the model's parameters, which enables targeted adjustments to the model without extensive retraining (Houlsby et al., 2019a). Another approach is prompt tuning, in which prompts get appended to the input data to guide the model's predictions in a task-specific manner. This method requires changes to only the prompt parameters rather than the model's core parameters (Lester et al., 2021; Jia et al., 2022). Finally, Low-Rank Adaptation (LoRA) focuses on altering the rank of weight matrices within the model to achieve efficient finetuning. This technique modifies specific components of the model's architecture to enhance performance on new tasks while minimizing the computational resources required for training (Hu et al., 2021).

2.5.4 AdaLora

AdaLoRA (Adaptive Low-Rank Adaptation) is a sophisticated PEFT technique. This method builds on (LoRA) and additionally incorporates features that enhance adaptability and optimize performance across diverse domains and datasets (Zhang et al., 2023).

The innovative aspect of AdaLoRA is its adaptability: It is able to dynamically adjust the parameters based on the task at hand. This adaptability is achieved through a mechanism that assesses the importance of specific model components in relation to the task.AdaLoRA then distributes parameters across weight matrices based on their significance scores. Specifically, AdaLoRA models the incremental updates using singular value decomposition (Park et al., 2024) to parameterize these changes. This approach allows for selective finetuning, targeting only the

most impactful parameters, thus reducing computational overhead and enhancing adaptability to new tasks.

AdaLoRA has demonstrated superior performance compared to other PEFT methods, such as traditional LoRA and full model finetuning, particularly in domains requiring high adaptability and where deployment constraints, such as computational resources, are a concern. This efficiency makes it an excellent choice for real-world applications (Zhang et al., 2023).

3 Related Work

In this chapter, we present recent advancements in ASR technology, specifically focusing on domain-adaptation for senior citizens and children using the Whisper model among others. The studies discussed here employ innovative methodologies to enhance ASR accuracy for these distinct demographic groups, contributing to the broader applicability and inclusiveness of ASR systems.

3.1 Domain Adaptation in the context of Senior Citizen German

Pekarek-Rosin and Wermter (2023) explore innovative methods to enhance ASR systems, particularly for the German language of senior citizens, by focusing on domain adaptation techniques. when it comes to specific languages and speaker demographics, there exist limitations to large-scale ASR models. The researchers seek to overcome this problem by improving the adaptability of these models to more specialized groups, such as German-speaking senior citizens.

Pekarek-Rosin and Wermter (2023) critically address the challenge of model performance degradation, known as catastrophic forgetting, when models trained on a broad domain are finetuned for specific, smaller domains. The authors propose a solution through selective layer freezing during finetuning and the incorporation of Experience Replay (ER) for continual learning. This approach aims at maintaining general speech recognition capabilities while enhancing performance on the target domain.

The experiments conducted involve finetuning state-of-the-art pre-trained multilingual speech recognition models, including the Whisper-small model, XLSR-53, and XLS-R. With XLSR-53 pre-trained on 53 languages and XLS-R on 128 (Conneau et al., 2021), these models enhance speech recognition performance across various languages by using similarities found during the training process (Babu et al., 2022). As discussed, Whisper is a recent large-scale multilingual model developed through unsupervised training. This model facilitates zero-shot cross-lingual speech recognition, speech translation, and language identification, spanning 97 languages and incorporating 680,000 hours of speech data (Radford et al., 2023). The models incorporate punctuation to varying extents; however, to ensure a fair comparison, the researchers normalized the produced transcripts prior to evaluation.

Pekarek-Rosin and Wermter (2023) finetuned these ASR models on the German Senior Voice Commands (SVC-de) dataset, which they designed themselves for the creation of an ASR system tailored for use within a home assistant framework. The dataset includes audio snippets dictated by a total of 30 German speakers, 21 of them were female, 9 were male, all of them aged between 50 and 99 years. The total duration of the recordings amounts to 3h 9m, equally split amongst the speakers. After recording, the sentences were manually segmented and transcribed to accurately assess the performance of ASR models under consideration. Although the majority of transcriptions closely aligned with the original utterances, variations were noted where some participants changed some of the sentences by repetition, omission, addition of words, or by rephrasing the sentences entirely. The second dataset used for finetuning was Common Voice DE (CV-de) (Ardila et al., 2020a), which is one of the most expansive and most frequently used German speech datasets for ASR. Since it was created within a crowd-source project, VD-de includes a diverse array of recording conditions and speaker profiles. The version used for this research, CV-de 10.0, includes 1,136 hours of verified audio contributed by 16,944 distinct speakers, and contains supplementary demographic information such as age group, gender, and accent of the speakers.

The researchers trained all the models with a batch size of 128 and the AdamW optimizer. They set the learning rate to 3e-4 for XLS-R and XLSR-53, and to 3e-5 for Whisper. By experimenting with different layer configurations and applying Experience Replay, the study assesses the impact on WERs for both the target domain (SVC-de) and the general domain. The findings demonstrate that the optimal model identified in this study is the pre-trained Whisper-small architecture, specifically tailored for German and refined on the SVC-de dataset utilizing 10% ER, achieving a reduction in WER from 18.4% to 3.0%. This improvement was primarily achieved by modifying the final six layers of both the encoder and decoder, which also helped to maintain CV-de's performance at a WER of 18.1%. Incorporating additional data from the original training domain further decreases the WER for the original domain. Nevertheless, it was noted that increasing ER to 20% introduces a compromise, enhancing CV-de's performance at the expense of the newly adapted domain's accuracy.

This work presents a significant contribution to the field of ASR by demonstrating the effectiveness of layer-specific finetuning combined with continual learning strategies to accommodate specific speaker demographics while preserving broad applicability. It opens avenues for future research on domain adaptation across various dialects and demographic groups, promising enhanced inclusivity and accessibility in ASR technology. The methods employed are model and dataset independent, suggesting their potential applicability to other domains and languages beyond German and the specific case of senior voice commands.

3.2 Domain Adaptation in the context of Child Speech

Jain et al. (2023) researched the potential of adapting large-scale ASR models, specifically the Whisper model and wav2vec 2.0, to improve recognition accuracy for child speech. This exploration is critical, given the inherent challenges in transcribing child speech, notably due to the scarce availability of comprehensive child speech datasets necessary for training child-centric ASR models effectively. Jain et al. (2023)) address the issue of a decline in performance of ASR systems when transitioning from adult to child speech. They highlight the distinct acoustic, linguistic, and phonetic characteristics that differentiate child speech from adult speech. Leveraging the multilingual and extensive dataset on which Whisper was trained, the study aims to determine the effectiveness of finetuning Whisper models with child speech data, thereby enhancing ASR performance for this demographic.

For finetuning, Jain et al. (2023) used the following child speech datasets: MyST Corpus (Pradhan et al., 2023), PFSTAR dataset (Russell, 2006), CMU Kids dataset (Eskenazi et al., 1997) and LibriTTS dev-clean dataset (Zen et al., 2019). All datasets underwent preprocessing to remove abbreviations, punctuation marks, white spaces, and any non-alphanumeric symbols, with all text converted to lowercase. Additionally, audio files were standardized to a 16kHz sampling rate and converted to 16-bit mono format. The dev-clean subset from the LibriTTS dataset, comprising 9 hours of audio, was utilized for evaluation. The My Science Tutor (MyST) Corpus, an American English dataset for child speech consisting of more than 393 hours, with 197 hours fully annotated, was segmented creating two portions: 55 hours for training purposes and 10 hours for evaluation. The PFSTAR dataset, encompassing recordings of British English-speaking children totaling 12 hours, allocated 10 hours for training and the remaining 2 hours for testing. The CMU Kids corpus was used exclusively for validation and comprises 9 hours of children's read-aloud sentences at Carnegie Mellon University. Even though these datasets are not as expansive as the Common Voice Datasets, for example, they represent the most comprehensive public resources for child speech research available to date.

In the finetuning process, a learning rate of 1e-5 was applied across all Whisper model finetuning experiments. The wav2vec2-base model underwent finetuning with a learning rate of 1e-4, and the wav2vec2-large model with a rate of 2.5e-5, aligning with established guidelines.The settings for finetuning were maintained as originally specified for both Whisper and wav2vec2 models: the strategy for both models included training only the models' final layers while keeping the rest unchanged, as recommended by their creators (Radford et al., 2023; Baevski et al., 2020). During the finetuning process, the Whisper model optimizes performance

by reducing the cross-entropy loss, in contrast, the wav2vec2 model aims to lower the Connectionist Temporal Classification (CTC) loss.

Employing this experimental setup, Jain et al. (2023) conducted two sets of experiments. In the first set, Whisper models of varying sizes (tiny, base, small, medium, large, and large-v2) were assessed across various child speech datasets. Notably, the large-v2 model underwent training for 2.5 times more epochs than the large model and incorporated additional parameters for enhanced regularization. Each model variant was available in two forms: one trained on multilingual datasets and another trained exclusively on English, except for the large and large-v2 models, which lacked an English-only version. A comparison of WER across English adult speech datasets revealed that models with a larger number of parameters typically achieved lower WERs, a trend also observed when these models were evaluated on child speech datasets. The second set of experiments focused on finetuning the Whisper models using the child speech datasets individually and combined. For this, the three models that exhibited superior performance in the initial trials (medium, medium.en, and large-v2) were used. Finetuning showed significant WER improvements. Particularly, finetuning on MyST 55h reduced WERs significantly in child speech tests but saw a 2% WER increase in the CMU test dataset. The Large V2 model demonstrated the most substantial WER reductions, with the lowest WER recorded at 2.88 on the PFS test. Across all finetuning experiments, models finetuned with datasets closely matching the test datasets' distribution showed the best results.

When comparing finetuned Whisper and wav2vec2 models, wav2vec2 models, previously finetuned with Librispeech, showcased superior performance over child speech datasets, except for the CMU test set where Whisper models performed better, indicating that the CMU Kids dataset might share acoustic properties with adult speech. Wav2vec2 models finetuned on MyST 55h specifically showed better WER rates compared to Whisper models on the same test, indicating a stronger generalization to similar dataset distributions. However, when combining MyST 55h and PFSTAR 10h for finetuning, wav2vec2 models still surpassed Whisper models across child speech datasets, suggesting a slightly better performance by wav2vec2 in child speech recognition. The findings reveal significant enhancements in ASR accuracy for child speech through finetuning and affirm the potential of these methodologies to elevate ASR technologies for children. Furthermore, the comparative analysis between Whisper and wav2vec2 models enriches the discourse on the efficacy of supervised versus self-supervised learning approaches in ASR.

3.3 German Automatic Speech Recognition Datasets

Wirth and Peinl (2023) designed a dataset compromising 610 hours of aligned audio-transcript pairs from German political debates, providing a robust resource for training ASR systems in recognizing formal German used in political contexts. While not medical, the formal language used is to some extent assimilable to the structured language found in medical transcripts. This makes the dataset a valuable resource for understanding high-context communication within German.

Rumberg et al. (2022) focused on German children's speech. Their corpus "kidsTALC" includes audio and phonetic transcriptions intended to support ASR technologies in educational and pediatric environments. Similar to medical ASR, recognizing child speech requires tailored models that can handle varied speech patterns and irregularities, which is crucial for applications that require understanding less structured language.

Within the Mozilla Common Voice project, Ardila et al. (2019) created an open and publicly reusable database of voice inputs from various languages, including German. The German segment of this dataset is particularly relevant for its broad demographic and linguistic diversity, offering a rich resource for training and finetuning ASR systems across different accents, dialects, and speech patterns. This dataset helps in developing more robust and inclusive ASR systems that can perform well across the diverse spectrum of German speech, which is essential for both general and specialized applications like medical ASR Eberhard and Zesch (2021).

While these datasets contribute significantly to the field of ASR by providing diverse linguistic and demographic data, they collectively underscore a persistent deficiency in domain-specific resources for the German medical sector, especially German medical datasets including Standard Austrian German pronunciation. Notably, there is a distinct lack of datasets that focus on Standard Austrian German, which features unique pronunciation and linguistic characteristics that are underrepresented in general German speech recognition models. This gap is particularly problematic in medical ASR applications, where accuracy and regional linguistic specificity are crucial for effective communication and information processing. The "MedASR-DE-AT" dataset we propose in this thesis is developed to fill this gap, offering in-domain audio data to enhance ASR applications within (Austrian Standard) German medical contexts.

4 Datasets

In this chapter, we explain the creation of the datasets used for our experiments.

Our main dataset, MedASR-DE-AT includes 4,480 audio recordings with corresponding transcriptions. Each recording is under 30 seconds in duration, totaling about 11.5 hours of audio. These recordings are derived from anonymized surgical patient records collected across German clinics from 2004 to 2016. The data was initially provided as a single large text file by the company 4voice AG. We then split the provided text file into approximately 90,000 separate files, each containing not more than one sentence to facilitate data handling and dictation.

4.1 Data Collection

Contrary to typical methodologies involving the transcription of existing audio (Simon et al., 2022), this project required creating new audio files from text due to the initial lack of audio data. The dataset comprises audio recordings dictated by two native Austrian Standard German speakers. Speaker 1, the primary contributor and author of this paper, dictated approximately 4,000 of the total 4,480 recordings. This speaker is female, was 25 years old at the time of recording, born in Vienna, and had prior experience in dictation, although lacking a medical background. Speaker 2, who contributed the remaining recordings, is a 39-year-old male from Lower Austria. Similar to Speaker 1, he had experience in dictation but did not possess a medical background.

4.2 Dictation Setting

Recordings were made using the Philips SpeechMike Premium [3]. Even though we had experience in dictating with this microphone and knew that it excels in handling background noise, we conducted the recordings in controlled, quiet environments to ensure clarity. Quality control was maintained by re-recording any audio that contained errors during dictation. Random audio samples were later verified against their transcripts to ensure accuracy.

[3] https://www.dictation.philips.com/de/produkte/stationaeres-diktieren/speechmike-premium-diktiermikrofon-lfh3500/

4.3 Data Preparation

We developed this dataset to serve as a tool for the further refinement and retraining of existing ASR models, thereby requiring compatibility with various model engines. Specifically, to ensure compatibility with the Whisper engine, it is essential that the audio files in this dataset do not exceed a duration of 30 seconds. This limitation is critical to ensure optimal performance of the system (Radford et al., 2023). Hence, we checked all audio files to ensure they adhered to the 30-second limit, discarding any that exceeded this duration to maintain consistency across the dataset. We decided to include medical notations such as numbers, punctuation, and abbreviations. These elements are essential in medical settings to convey accurate information: subtle distinctions, such as precise dosage amounts or specific medical shorthand, may substantially modify the meaning. The inclusion of these critical elements in datasets is crucial for the training of realistic ASR systems, as it ensures a reliable interpretation of complex medical language and abbreviations prevalent in healthcare settings. Furthermore, this inclusion aids in minimizing the risk of miscommunication and enhances the overall efficacy of technological applications in patient care. It can mitigate potential errors and improve the quality of information exchange in healthcare environments, consequently contributing to increased patient safety (Pakoci et al., 2022).

4.4 Statistical Analysis of Training Data

We divided the MedASR-DE-AT dataset into three sections: 80% for training, 10% for validation, and 10% for testing. Specifically, the training set contains 3,584 instances totaling 8 hours, 54 minutes, and 47 seconds. The validation set includes 449 instances amounting to 1 hours, 9 minutes, and 4 seconds, and the testing set comprises 448 instances, totaling 1 hours, 6 minutes, and 40 seconds.

To enhance our understanding of the training data, we conducted a thorough analysis of the textual content within the train split of our dataset. This examination focuses on vocabulary size, sentence structure, and prevailing linguistic patterns, offering a comprehensive description of the dataset's linguistic features.

The training data compromises a total word count of 43,686 and a vocabulary size of 7,747 unique words. This indicates a moderate level of lexical diversity, suggesting that the dataset encompasses a wide range of language use, which is crucial for comprehensive language modeling and understanding. The ratio of total words to unique words provides a measure of lexical richness, which in this case suggests a relatively diverse usage of language but with some repetition of vocabulary, which seems logical for domain-specific datasets like ours.

Figure 2 illustrates the distribution of sentence lengths within the training data.

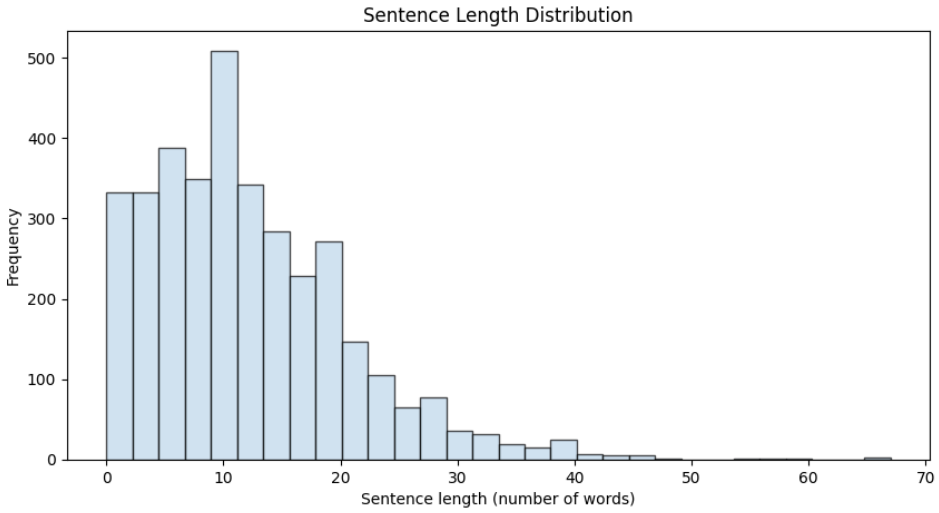

Figure 2: Distribution of Sentence Length of Training Data

The average sentence lengths is approximately 12 words. Most of the training sentences compromise between 10 to 20 words. The presence of sentences with zero length results from entries composed solely of numbers and punctuation, which have been removed during preprocessing.

Further, we identified the most common words and bigrams in the training data. The results are presented in Figure 3 and Figure 4, respectively. These frequent terms reflect specific themes or focus areas within the dataset. For instance, the prevalence of directional terms (*rechts*, *links* meaning *right*, *left*) and procedural terms (*Operation*, *einbringen* meaning *operation*, *place*: like in *place the catheter*) strongly indicate the medical context.

The analysis of bigrams further underscores the specific vocabulary used but also highlight typical actions or procedures described in the dataset.

Additionally, we conducted an analysis of the audio files featured in the training data. As mentioned, the total playtime of all audio files in the training set amounts to 8 hours, 54 minutes, and 47 seconds, indicating a substantial amount of audio data for processing and analysis. The average duration of an audio file is approximately 8.95 seconds, and the median duration is 7.89 seconds. Since the Whisper engine expects audio snippets to be under 30 seconds long for optimal processing, we did not include any files longer than 30 seconds in our dataset. The shortest audio file in the dataset is just 0.23 seconds, whereas the longest extends to 29.81 seconds. Further, there is a standard deviation of 4.86 seconds, which indicates a moderate dispersion around the average length. This variability shows that while many audio files are close to the average length, there are also numerous outliers, contributing

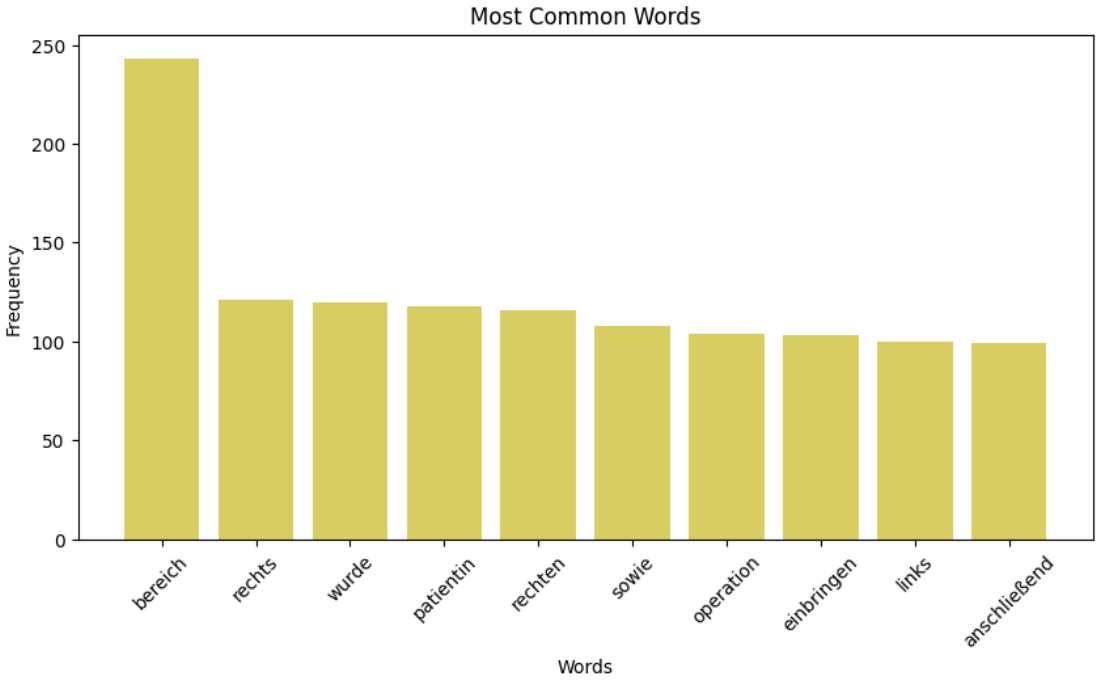

Figure 3: Most Common Words in Training Data

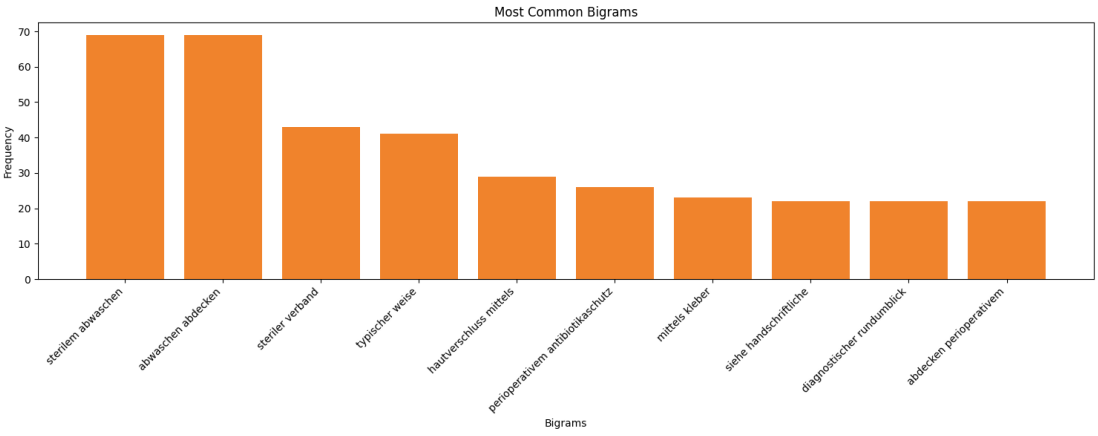

Figure 4: Most Common Bigrams in Training Data

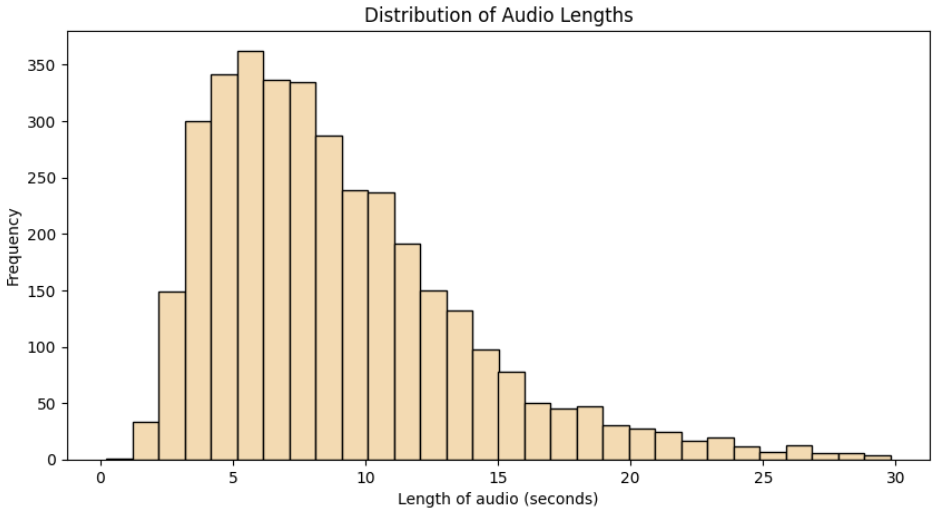

Figure 5: Distribution of Audio Lengths in Training Data

to the wide range in audio file lengths that can be observed in Figure 5.

4.5 Dataset Structure

The dataset is modeled after widely used audio datasets like the Common Voice speech corpus (Ardila et al., 2019) and organized according to best practices for creating audio datasets from Hugging Face [4]. We designed the dataset following the Hugging Face AudioFolder structure.

The data folder can be named arbitrarily and should hold the audio files, which can be in either wav or mp3 format; for this dataset, we used the wav format. The metadata file should be either a csv or a json file and needs to have a column labeled file_name that associates each audio file with its corresponding metadata. For our purposes, a csv file format has been utilized. Table 1 shows how the metadata file is composed.

4.5.1 Separate Dataset for Testing on Standard German

As mentioned in Section 4.4, the MedASR-DE-AT dataset was partitioned into three subsets: training (80%), validation (10%), and testing (10%). However, we additionally created a separate dataset for testing on Standard German pronunciation, named MedASR-DE-test on Hugging Face, including 158 audio recordings alongside

[4] https://huggingface.co/docs/datasets/audio_dataset

Table 1: Example of medical data transcriptions.

file_name	**text**
data/audio_00001.wav	Diagnosen: Obere gastrointestinale Blutung bei Ösophagusvarizen, portale Dekompensation der Leberzirrhose mit Ösophagusvarizen; Fundusvarizen; Varizenknäuel unterhalb Milzhilus; Analvarizen; keine Dünndarmvarizen.
data/audio_00002.wav	08/03 Erstdiagnose von Ösophagusvarizen, jetzt Beginn Ligaturbehandlung.
data/audio_00003.wav	Ambulant erworbene Pneumonie, großvolumiges, kompaktes Oberlappeninfiltrat rechts.

their corresponding textual transcriptions. Like the MedASR-DE-AT set, these materials were sourced from the same textual data provided by 4voiceAG, ensuring consistency in data quality and context. The audio recordings were produced by a female speaker native to Germany, utilizing Standard German and differentiating her linguistic background from ASG. This distinction is significant as it adds an additional layer to the evaluation process: the potential impact of training models with Austrian Standard German pronunciation and testing them on Standard German pronunciation. This setup provides a unique opportunity to observe any effects or discrepancies that may arise from the linguistic differences between these two variants.

4.6 Limitations

The dataset is comprised of audio recordings from two speakers, with the majority provided by one individual, which may limit variation in speaker characteristics. Both speakers do not have a medical background, which might affect the pronunciation accuracy of medical terms. Additionally, the content is sourced exclusively from surgical patient records, focusing the dataset's applicability primarily within the surgical domain rather than the broader medical field. Both datasets are private and can not be made open source.

4.7 Ethical Considerations

All personal data was anonymized by the data provider to protect patient privacy. This de-identification and approval from the data owners to show data snippets in the thesis ensure the ethical use of this data.

4.8 Possible Applications

The MedASR-DE-AT dataset not only fills an essential gap in ASR resources but also opens up numerous possibilities for research and application.

In our project, we use the dataset to train and finetune Whisper models. However, researchers can also finetune other pre-trained SOTA ASR models like Wav2Vec 2.0 (Hsu et al., 2021), or BERT-based speech models (Devlin et al., 2019) specifically for the German-speaking medical domain, potentially reducing the WER and improving recognition accuracy for medical terminologies. For finetuning the Whisper model, we suggest following the Hugging Face tutorial: finetune Whisper For Multilingual ASR with HF Transformers [5]. Lowering the WER is crucial for applications where precision of information is critical such as medical prescriptions or patient records (Errattahi et al., 2015a).

Furthermore, this dataset provides an opportunity to study the impact of training models on Austrian German data and testing on Standard German, offering insights into linguistic adaptation and model robustness. Austrian German and Standard German, while largely mutually intelligible, contain distinct phonetic, lexical, and grammatical differences that could potentially affect the performance of ASR models (Wirth and Peinl, 2023). Training models on such region-specific datasets allows researchers to study the adaptability and flexibility of these models when applied to a linguistically different test set. Furthermore, by evaluating how models trained on Austrian German adapt to Standard German, researchers can gain valuable insights into the linguistic adaptation capabilities of current machine learning algorithms. This involves understanding how well a model can transfer learned knowledge from one dialect to another and identifying which model architectures are more robust against linguistic variations. Such studies are important for developing ASR systems that are effective across different dialects and accents in order to improve user experience and system reliability in real-world applications.

Finally, practitioners can use the dataset to create specialized ASR systems that are better equipped to handle the unique challenges of medical transcription. These systems can be optimized to accurately interpret the complex and specialized terminology and acronyms frequently used in the medical field. By enhancing the capability of ASR systems to recognize and process such specialized language, the utility of ASR technology in healthcare settings can be significantly enhanced. This leads to improvements in the efficiency of documenting medical records and assists in facilitating the administrative workflows within healthcare facilities, thereby supporting clinical staff by allowing them more time to focus on patient care rather than paperwork (Enarvi et al., 2020).

This dataset is an important step towards improving German medical ASR

[5] https://huggingface.co/blog/finetune-whisper

systems. By addressing specific linguistic needs and providing high-quality, relevant training data, this contributes to the development of more accurate and efficient ASR technology in medical settings where precision is vital. A notable distinction of this dataset is its focus on Austrian Standard German, a language variant that, until now, has not been particularly catered for in existing resources. Typically, datasets have concentrated on the Standard German language, thereby overlooking the nuanced differences presented by the Austrian variation. With this unique dataset, we not only fill a critical gap within medical ASR but also broaden the applicability and effectiveness of ASR systems in regions where Austrian German is prevalent. In the next step of our research, we will focus on finetuning Whisper models with this dataset to determine if targeted retraining and parameter adjustments can improve ASR accuracy, specifically for Austrian Standard German in medical settings.

5 Baseline

In the following section, we first present the model selected for the domain adaptation process and discuss the considerations made to ensure that the model not only aligns with the research objectives but also adequately addresses the complexities of the data involved.

5.1 OpenAI's Whisper Model

As discussed in Section 2.4.3, OpenAI's SOTA ASR model Whisper was trained on an extensive dataset of 680,000 hours of multitask audio comprising English transcriptions, translations from various languages to English, non-English transcriptions, and pure instrumental audio sourced from the internet. Hence, its architecture is adept at processing a broad spectrum of speech inputs, from transcriptions in multiple languages to segments without speech, only containing instrumental inputs. Furthermore, the model is able to perform a multitude of speech processing tasks concurrently, which is crucial for its ability to adapt to specialized domains such as medical speech recognition (Radford et al., 2023). Figure 6 illustrates the model's operational framework and training methodology.

Central to Whisper's design is a Transformer-based neural network structure, employing a sequence-to-sequence learning mechanism. This setup begins with the transformation of audio inputs into an 80-channel, or 128-channel for newer versions, log-magnitude Mel spectrogram. The representation then undergoes a convolutional process, augmented with a GELU activation function. The Gaussian Error Linear Unit (GELU) activation function is a type of activation function used in deep learning models. It has become increasingly popular for its ability to improve learning capacity, stability, and computational efficiency of models. GELU is particularly noted for its differentiability, boundedness, stationarity, and smoothness, making it superior to traditional activation functions like the Rectified Linear Unit (ReLU) in various applications (Lee, 2023). After the convolutional process, the processed data is being passed through successive Transformer encoder blocks. The encoder's output feeds into Transformer decoder blocks, where a sequence of tokens correlating with the transcribed text is generated.

A complete speech processing system comprises various components beyond speech recognition, such as voice activity detection, speaker diarization, and in-

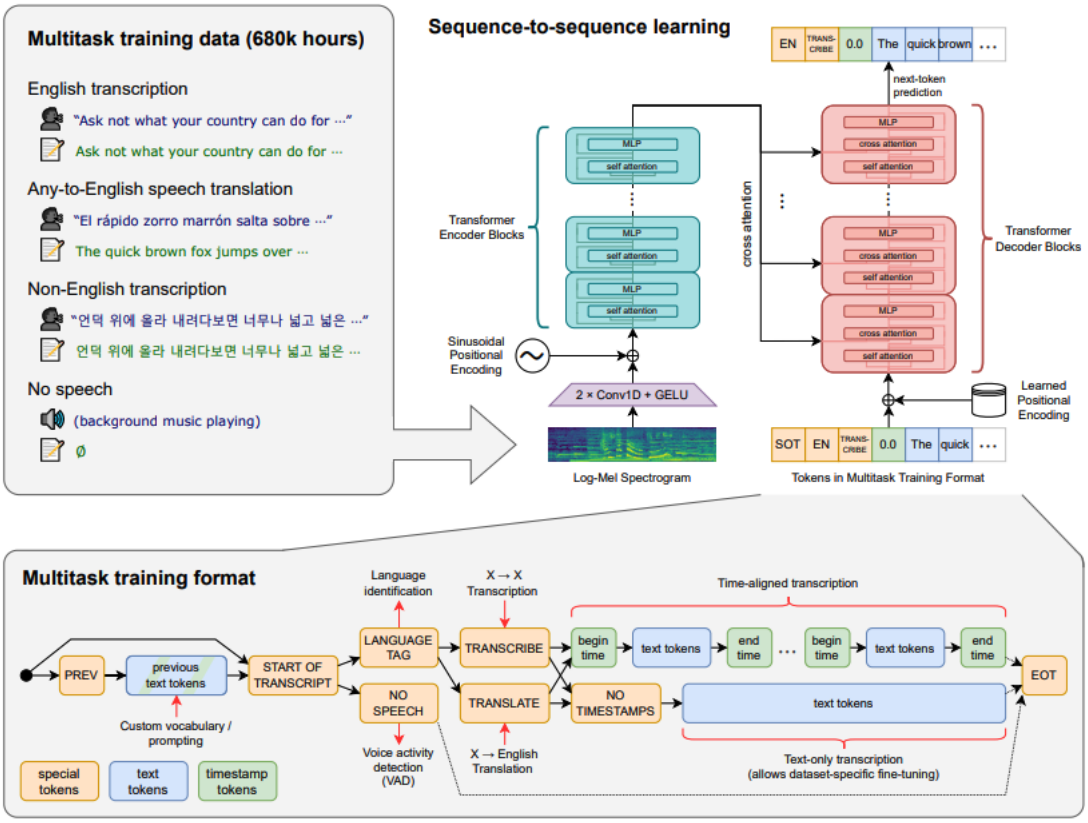

Figure 6: Overview of the Whisper Model. Figure by Radford et al. (2023)

verse text normalization. These elements are typically managed through distinct pipelines, creating a complex system with multiple interacting parts. With the Whisper model, Radford et al. (2023) propose a unified model that handles the entire processing pipeline. The model discerns task-specific nuances by employing tokens that represent actions such as transcription, translation, voice activity detection, alignment, and language identification. The researchers define all tasks and conditioning information through a sequence of input tokens for the decoder. The model is also trained to utilize the history of the transcript text to enhance its ability to resolve ambiguities in the audio and occasionally incorporates preceding transcript text into the decoder's context.

Task initiation and conclusion are marked by special tokens indicating the start and end of a transcript, ensuring the model accurately captures the context within each training segment: prediction begins with a specific token, followed by identifying the language spoken, indicated by a unique token for each language in the 99-language training set derived from the VoxLingua107 model. If an audio

segment contains no speech, the model predicts a silence token. Subsequently, a token designates the task, either transcribing or another specified task. The decision to predict timestamps is made by including a particular token, defining the task and format before beginning the output of tokens. Timestamp predictions are made relative to the audio segment, quantized to the nearest 20 milliseconds, aligning with the Whisper model's time resolution. Corresponding tokens have been added to the model's vocabulary and interleave these predictions with the text captions, predicting start times before and end times after each caption's text. For transcript segments partially included in a 30-second audio chunk, the model predicts only the start time, which cues the model to align subsequent decoding with that timestamp; otherwise, the audio is truncated to exclude the segment. Finally, a closure token is added. The training loss is masked over prior context text and used to train the model to predict all other tokens (Radford et al., 2023).

Radford et al. (2023) assessed the performance of long-form transcription on seven diverse datasets composed of speech recordings that vary in length and recording conditions. These datasets were specifically selected to encompass a wide range of data distributions. They include a long-form version of TEDLIUM3 (Hernandez et al., 2018), which has been adapted to feature each TED talk in its entirety. Additionally, they examined jargon-heavy segments from The Late Show with Stephen Colbert (Meanwhile), various sets of videos and podcasts commonly utilized as ASR benchmarks in online blogs (Rev16 and Kincaid46), recordings from financial earnings calls (Rio et al., 2021), and comprehensive interviews from the Corpus of Regional African American Language (CORAAL) (Gunter et al., 2021). Radford et al. (2023) research shows that Whisper consistently outperforms the compared models and that this is particularly notable on the Meanwhile dataset, which is characterized by a high incidence of uncommon words. Whisper's enhanced performance on such datasets underscores its robustness and advanced capability to handle diverse linguistic elements and it further establishes its superiority over competing ASR technologies. This proficiency is essential for applications requiring accurate interpretation and transcription of complex and specialized vocabulary, such as the medical domain, where precision and adaptability to complex terminologies are paramount (Chang et al., 2023).

Radford et al. (2023) also reveal that, particularly for English ASR tasks, the system's performance is nearing the precision typically attributed to human transcriptionists, which is considered the gold standard in the field of speech recognition (Gref et al., 2022). This close approximation to human-level accuracy indicates significant advancements in the underlying algorithms and model training methodologies employed by Whisper. These improvements in performance are critical as they enhance the usability and effectiveness of ASR technologies across various applications, including real-time communication, automated transcription services,

and assistive technologies, also in the healthcare domain (Zhou et al., 2018).

5.1.1 Selection Considerations

We selected the Whisper architecture for our project primarily due to its robust performance in zero-shot and out of distribution settings, making it highly adaptable for domain-specific tasks like medical German ASR (Peng et al., 2023). Unlike other models that may require extensive finetuning on vast amounts of domain-specific data, Whisper's large-scale weakly supervised training allows it to adapt to new audio distributions effectively. The Whisper model is remarkably efficient at zero-shot learning and can generalize across different languages and dialects without the need for dataset-specific finetuning. This capability makes it an ideal choice for applications requiring high adaptability and reliability in speech recognition, such as in the medical sector, where accuracy and versatility are crucial.

Moreover, Whisper's effectiveness in handling multiple languages and its proven capability in domain adaptation make it especially suitable for the complex terminology and syntax of medical German (Chang et al., 2023). Its multitask learning format, which includes speech recognition alongside translation and other language tasks, enables the model to better understand and process the specialized vocabulary and sentence structures typical in medical contexts. This adaptability is enhanced by the model's architecture, which avoids the fragmentation of tokens in languages other than English, ensuring more coherent and accurate recognition of medical terms.

Beyond its adaptability and performance, the decision to utilize the Whisper model was also influenced by factors such as the availability of its training code and pre-trained models, active community support, and comprehensive documentation. These aspects ensure ease of use and ongoing improvements through community contributions, which are invaluable for a research-oriented project.

5.2 Model Versions

We use two different versions of the Whisper model released by OpenAI as baselines: Whisper-medium (769 million parameters) and Whisper large-v3 (1550 million parameters) (Radford et al., 2023). The medium model was trained on either English-only or multilingual data and Whisper-large-v3 was trained on multilingual only. The Whisper large-v3 model has the same foundational architecture as its predecessors with specific adjustments. While the previous versions input audios with 80-channel log-magnitude Mel spectrograms, Whisper-large-v3 incorporates an expanded input of 128 Mel frequency bins, signifying that the model is capable of capturing more detailed audio information. The large-v3 model was trained

on 1 million hours of weakly labeled audio and an additional 4 million hours of pseudolabeled audio, sourced from the Whisper large-v2 model, for 2.0 epochs. Notably, the large-v3 model demonstrates enhanced performance across various languages, with error rates showing reductions of up to 20% compared to the Whisper large-v2 model in general scenarios [6].

[6] https://huggingface.co/openai/whisper-large-v3

6 Experiments

In this following section, we will present the experimental design implemented throughout this thesis.

6.1 Benchmark of Original Whisper Models

In ASR, three error types occur: substitutions (S), deletions (D) and insertions (I). Substitutions happen when a reference word is transcribed incorrectly as another word. Deletions are omissions of words in the generated transcriptions, and Insertions occur when a word appears in the generated transcription without a corresponding word in the reference sequence. A fundamental challenge in calculating the performance for ASR systems is establishing word alignment between the reference and the automatic transcription. This alignment process is crucial for identifying which words have been deleted, inserted, or substituted (Errattahi et al., 2015b). The most commonly used metric for evaluating the performance of ASR systems is the WER (Ali and Renals, 2020). This metric provides a standard for comparing different systems or methods and quantifies their performance and progress in specific tasks through the analysis of error statistics: it quantifies the proportion of incorrect words in the generated transcriptions, notably substitutions, deletions, and insertions, relative to the total number of words present in the source transcriptions (Errattahi et al., 2015b). The WER is defined as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{H + S + D} \tag{1}$$

where S, D, and I denote the number or substitutions, deletions and insertions, respectively, and N is the total number of words in the source transcriptions. This equation can be specified further, when dissecting N into its constituent parts: H, S, and D, where H is the amount of hits, in other words the number of correct words. Hence, the WER takes on a percentage, typically between 0 and 1, with low percentages indicating a low error level and higher percentages showing a higher error level. For instance, a WER of 0.1, or 10%, signifies that 10% of the output words are wrong, implying that 90% of the input words were transcribed accurately. It is important to note, however, that the WER may exceed 1. This occurs when the number of errors (S+D+I) in the generated text surpasses the number of words

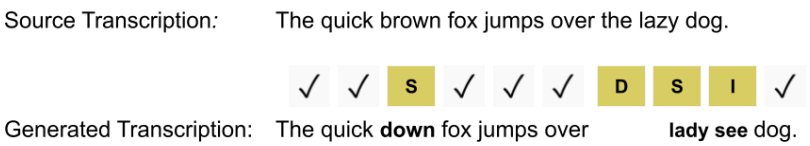

Figure 7: Word Error Rate Explanation

in the reference text (Morris et al., 2004). Figure 7 shows an example for better understanding.

In the case of the example in Figure 7, the WER would be computed as shown in the equation 2. The result is an error rate of approximately 44%, indicating a rather high error level, where nearly half of the words have been transcribed incorrectly.

$$WER = \frac{S + D + I}{N} = \frac{2 + 1 + 1}{9} = 0,44 \tag{2}$$

To evaluate the efficacy of performing domain adaptation and finetuning methods on Whisper models, and to compare them to our finetuned models, we established performance benchmarks using the original models. We performed the evaluation on the automatic transcription of audio recordings in the test sets from MedASR-DE-AT and the MedASR-DE-test dataset. The first one, the test split from our dataset MedASR-DE-AT, comprises 448 audio recordings done by the two speakers with Austrian Standard German pronunciation. The second, MedASR-DE-test, includes 158 audio recordings articulated by the speaker with German Standard pronunciation.

Evaluation For the computation, we used the evaluation script proposed by OpenAI; it is the same for all model versions [7]. The script employs the Python library evaluate [8], which is capable of assessing the performance of machine learning models and datasets. This library enables access to a wide range of evaluation methodologies, such as accuracy, BLEU score (Papineni et al., 2002), COMET (Rei et al., 2020) and WER, among other common metrics utilized in the field of NLP. By loading and employing the WER metric, the library compares predicted and reference transcriptions against each other and automatically carries out the WER calculation as discussed above. We opted to conduct two distinct evaluation phases: in the initial phase we we implement normalization steps as

[7] https://huggingface.co/openai/whisper-large-v2
[8] https://github.com/huggingface/evaluate

recommended by OpenAI [9] and carried out in related research projects (Jain et al., 2023; Pekarek-Rosin and Wermter, 2023), to ensure a more equitable evaluation of the model. For normalization, we use the Whisper class BasicTextNormalizer [10]. The BasicTextNormalizer applies several normalization steps to ensure the model's uniformity and improve recognition accuracy. It converts all text to lowercase to eliminate variations caused by capitalization. Additionally, it removes words enclosed within brackets and parentheses, which are often not crucial for the main analytical tasks. It also clears away special symbols and diacritics to rid the text of unnecessary characters. Finally, it reduces multiple whitespace characters to a single space, thereby standardizing the spacing within the text. These modifications contribute to the consistency of the text input, which is crucial for an effective computational analysis and optimal model performance. We performed all evaluations on Google Colab using a T4 GPU to ensure efficient processing and evaluation. The results of this initial evaluation phase are illustrated in Table 5.

6.2 Domain Adpatation: Finetuning Experiments

In the following section, we present the domain adaptation experiments employing finetuning methods conducted in this thesis. For all finetuning experiments, we implemented 4-bit quantization [11] to decrease the model's memory requirements and accelerate the processing speed. 4-bit quantization is a process used in machine learning that reduces the precision of the numbers representing model weights from their original format, which is typically 32-bit floating point, to just 4 bits. This significantly decreases the model's memory usage and speeds up computation by using less data to represent each weight (Weng, 2021).

6.2.1 Finetuning Whisper-medium

For our experiments finetuning the Whisper medium model for German medical speech data, we identified optimal settings through empirical testing. The model was trained using the Seq2SeqTrainer from Transformers [12] over three epochs with evaluations every 300 steps. We set the learning rate at 1e-5, with a training batch size of 8 and an evaluation batch size of 4. A warm-up ratio of 0.03 was used at the start of training, which translates to 90 warm-up steps given the total of 3000 steps across three epochs. The settings are displayed in Table 2.

[9] https://huggingface.co/openai/whisper-large-v2
[10] https://github.com/openai/whisper/blob/main/whisper/normalizers/basic.py
[11] https://huggingface.co/docs/transformers/main_classes/quantization
[12] https://huggingface.co/docs/transformers

Table 2: Training Settings of Finetuning Whisper-medium

	Parameters
Epochs	3
Steps per Epoch	1000
Evaluation steps	300
Total steps	3000
Learning Rate	1e-5
Training Batch Size	8
Evaluation Batch Size	4
Warmup Steps	90
Optimizer	AdamW
Loss Function	Cross Entropy Loss
GPU Model	NVIDIA L4
GPU Memory	25 GB

For optimization, the default optimizer of the Seq2SeqTrainer, the AdamW optimizer [13], is used to enhance learning dynamics, improve regularization, and stabilize updates. The loss function utilized for finetuning the Whisper model is the same as for the model itself: cross-entropy loss [14].

Training was conducted on Google Colab's NVIDIA L4 GPU [15], which has 25 GB of GPU memory, sufficient to handle the Whisper medium model's memory requirements, which often exceed 23 GB. This configuration was chosen over the less capable NVIDIA T4 GPU, which offers only 15 GB.

6.2.2 Impact of Training Data Volume on Model Performance

Additionally, we conducted an experiment to evaluate the impact of training data volume on the Whisper medium model's performance. The experiment involved training the model with varying amounts of data: 1 hour, 2 hours, 3 hours, 4 hours, 5 hours, 6 hours, 7 hours, 8 hours, and finally 9 hours, which compromises the entire training set. Each round of training was conducted with the same settings and training parameters: we trained for 1000 steps with evaluation at each 200 steps. The learning rate was set to 1e-5, the training batch size was 6, the evaluation batch size 4, and the warm-up steps were set to 0.03, which accords to 30 steps when having 1000 training steps. The smaller batch size and number

[13] https://huggingface.co/docs/transformers/v4.25.1/en/main_classes/trainer
[14] https://huggingface.co/sanchit-gandhi/whisper-small-hi/discussions/12
[15] https://cloud.google.com/compute/docs/gpus?hl=de

of training steps were necessary because of computational resources constraints. Despite having access to Colab Pro and purchasing additional compute units, it was essential to economize the use of training instances to manage resource availability effectively. Therefore, these limitations were strategically chosen to optimize the training process under the given constraints. Once again, the AdamW optimizer and cross-entropy loss have been employed and the experiments were conducted using the NVIDIA L4 GPU. The training parameters for training the Whisper-medium models with different amounts of training data are presented in Table 3.

Table 3: Training Settings of Finetuning Whisper-medium with different Amounts of Training Hours

	Parameters
Epochs	1
Steps per Epoch	1000
Evaluation steps	200
Total steps	1000
Learning Rate	1e-5
Training Batch Size	6
Evaluation Batch Size	4
Warmup Steps	30
Optimizer	AdamW
Loss Function	Cross Entropy Loss
GPU Model	NVIDIA L4
GPU Memory	25 GB

Two training conditions were tested: with normalization using the BasicTextNormalizer, as explained in Section 6.1, and without normalization. Subsequently, we conducted the same experiment on the MedASR-DE-test dataset, which includes solely audios with a Standard German pronunciation. The results are shown in Figure 15 and Figure 16.

6.2.3 Finetuning Whisper-large-v3

For our finetuning project using the Whisper-large-v3 model, we utilized AdaLoRA to accommodate the training of a large-scale model within constrained GPU memory environments, as discussed in Section 2.5.4. Training the full Whisper-large-v3 model on Google Colab was not possible, due to immediate CUDA memory overflow. AdaLora, however, reduced the GPU memory footprint to 10-11 GB during training.

We determined the best AdaLoRA configuration through practical experimentation. Specifically, the configuration included the following parameters: an initial rank (init_r) of 12, which is the starting complexity of the model adaptations, and a target rank (target_r) of 4, indicating the reduced computational demand by the end of training. The coefficients beta1 (0.9) and beta2 (0.99) were used for computing running averages of the gradient and its square, similar to those in the Adam optimizer. The training steps at which the rank reduction begins and completes were set at 200 (tinit) and 1000 (tfinal), respectively, with a deltaT of 10 indicating the interval of training steps after which adjustments in model complexity are applied. The configuration also specified a lora_alpha of 32, a scale factor for the learning rate applied specifically to the low-rank adapted parameters, and a lora_dropout of 0.03 to prevent overfitting. Additionally, an orth_reg_weight of 0.8 was set to encourage orthogonality in the adapted parameters, enhancing model stability and performance.

Table 4: AdaLora Training Configuration

	Parameters
init_r	12
target_r	4
beta1	0.9
beta2	0.99
tinit	1e-5
tfinal	200
deltaT	1000
lora_alpha	32
lora_dropout	0.03
orth_reg_weight	0.8

The training setup was the same as the one for the Whisper-medium model, involving a training batch size of 8, an evaluation batch size of 4, a learning rate of 1e-5, and warm-up steps set at 0.03, corresponding to 30 steps in a total of 3000 planned steps, aiming for 3 epochs, with evaluation at every 300 steps.

This configuration was strategically chosen to efficiently manage computational resources while ensuring effective model training dynamics.

6.3 Error Annotation

Since we wanted to have a deeper understanding of the models' performance, we tasked the Whisper models to generate transcriptions for audio samples from

our distinct datasets. The Whisper medium model completed the automatic transcription in approximately 2 hours, while the Whisper large-v3 model required over 4 hours, using the Google Colab CPU. This significant difference in processing time emphasizes the impact of model complexity, measured by the number of parameters, on computational effort and runtime. We then compared the generated transcriptions against the reference transcriptions.

For the analysis, we utilized Label Studio. We initially configured the platform with a label template tailored for Named Entity Recognition but modified the labels to suit our specific needs: MISR (misrecognitions), defined as a substitution of a non-medical term; OMI (omission); INS (insertion); MEDTERM (misrecognition or false writing of medical term); ORTH (orthographic errors); GRAMM (grammatical errors); NUM (numbers); and NOT (differences in notation).

To attain these labels, we followed the DQF-MQM TAUS Error Typology (Valli, 2015) format, but customized it to correspond to transcription-related errors only, similarly to Papadopoulou et al. (2021). Multidimensional Quality Metrics (MQM) is a framework used to evaluate and quantify the quality of translated texts by assessing various error types and their impact on overall translation accuracy and fluency (Freitag et al., 2021). In the context of transcription quality, MQM is equally useful because it allows for a detailed, structured assessment of transcription errors. This is done by including a categorization by type and severity, which helps in identifying specific areas for improvement. This systematic approach ensures that transcriptions maintain a high level of accuracy and adhere closely to the source texts, which is essential for effective communication and documentation (Choi, 2023).

We uploaded both the generated and reference transcriptions in a manner that displayed them one above the other, allowing for quick identification of discrepancies. This setup was applied for analyses of the Whisper-medium model, both in its original and finetuned forms. The transcriptions were created on our two datasets: MedASR-DE-AT, which utilizes Austrian Standard German pronunciation, and MedASR-DE-test, which features standard German pronunciation. This approach allows for a comparative evaluation of the model's performance across different dialectal variations within the German language, while assessing the impact of finetuning on transcription accuracy in medical speech recognition tasks. The results are presented in Section 7.3.

7 Results

In this chapter, we will present the results achieved in our benchmarking project, the finetuning experiments and the error annotation through Label Studio.

7.1 Benchmarking Original Whisper Model

Table 5: WER Results (%) ↓ of original Whisper models (with normalization)

	WER ↓ MedASR-DE-AT	**WER ↓ MedASR-DE-test**
Whisper-small	23.95	17.97
Whisper-medium	17.97	11.47
Whisper-large-v2	15.20	8.29
Whisper-large-v3	13.62	7.23

The data presented in Table 5 and 6 provides a clear insight into the impact of normalization on ASR performance. Normalization evidently lowers WER values across all Whisper model sizes, from small to large-v3. For instance, the WER for the Whisper-small model shows a reduction from 30.69% to 23.95% on the test set of the MedASR-DE-AT dataset when normalization is applied. Similar improvements can be seen for the Whisper-large-v3 model, where WER decreases from 20.88% to 13.62% on the same dataset. On the MedASR-DE-test set, the WER for the Whisper-large-v3 model was reduced to below 10%. This indicates a significant enhancement of model performance due to normalization. Yet, while normalization is beneficial, it is imperative for ongoing development to focus on enhancing the robustness of ASR models in handling raw, not normalized speech to ensure reliability in complex medical environments. Refinement of the acoustic and language modeling capabilities in these models may enhance their ability to accurately reflect the intricacies of medical speech and reduce their dependence on text normalization techniques.

Generally, the higher-version Whisper models, such as Whisper-large-v3 achieve better WER results. This observation is evident, since this model is more complex with a greater number of parameters and Mel frequency bins than its predecessors. This complexity enables it to capture nuanced aspects of language, including the

Table 6: WER Results (%) ↓ of original Whisper models (without normalization)

	WER ↓ MedASR-DE-AT	**WER ↓ MedASR-DE-test**
Whisper-small	30.69	22.23
Whisper-medium	24.89	15.61
Whisper-large-v2	22.20	13.19
Whisper-large-v3	20.88	11.67

specialized vocabulary and syntax used in medical documentation, more effectively than the smaller models.

Additionally, the WER values reveal some interesting results: both WER values are higher for MedASR-DE-AT, which was recorded by Austrian Standard German speakers. These findings indicate that the Whisper models perform significantly worse on Austrian Standard German pronunciation, as compared to Standard German. This discrepancy probably arises because the model is more used to Standard German pronunciation. Whisper has been trained on data from a variety of online sources which are generally public or available under licenses that permit academic and research use. Radford et al. (2023) do not provide a detailed breakdown of each specific source or the exact composition of the data in terms of languages or variants. However, it is likely that the majority of the German training instances consist of audio recordings from speakers of Standard German, given that the population of Standard German speakers is significantly larger than of Austrian Standard German speakers.

Moreover, Whisper performs notably better when tasked with transcribing everyday language: According to Radford et al. (2023) the Whisper-small model achieves a WER of 13%, while the Whisper-large-3 model records a WER of 6.4% on the CommonVoice 9.0 dataset (Ardila et al., 2020b), including 14,973 hours of vernacular in 93 languages. The persistence of relatively high error rates even after normalization, ranging from 7.23% to 30.69% across the models and datasets, highlights underlying limitations in the models' ability to accurately transcribe medical speech. These findings underscore that the highly complex medical terminology and its subtleties discussed in Section 1.2.1 and Section 1.3 do have an important impact on the model's performance.

7.2 Finetuning Results

Our primary objective of this research was to enhance the performance of the Whisper model in accurately transcribing audio recordings within the German medical domain. This goal was successfully achieved through systematic domain

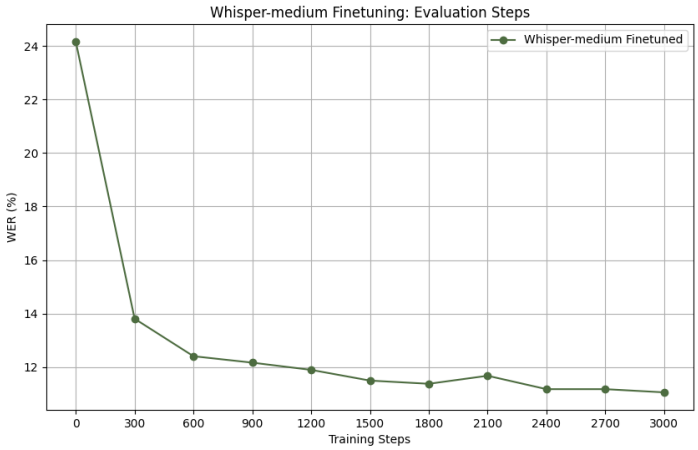

Figure 8: Whisper-medium Finetuning: Evaluation Steps

adaptation and finetuning approaches. By adapting the Whisper model to better recognize and process the specific terminologies and pronunciations characteristic of Austrian Standard German medical speech, we were able to significantly improve its transcription accuracy.

7.2.1 Finetuning Whisper-medium

Figure 8 illustrates the WER progression of the Whisper-medium model across evaluation steps during training. It is important to note that these values indicate the WER on the validation set of MedASR-DE-AT during training. The test and evaluation set are not identical, which accounts for some variation in error rates between the two sets.

Initially, the WER drops drastically from 24.15% to 13.81% after 300 training steps, demonstrating that the model quickly adjusted to the fundamental characteristics of the medical speech data. The decline in WER continues, but at a slower rate, reaching 12.41% at 600 steps and 11.38% by 1800 steps. This suggests that the model progressively refined its understanding of the data, improving its transcription performance. After 1800 steps, the WER fluctuates slightly, peaking with 11.68% at 2100 steps before declining again to stabilize around 11.18% at 2400 and 2700 steps, and finally reducing slightly to 11.06% at 3000 steps. These fluctuations might be attributed to the model's adjustments to optimize performance on complex aspects of the evaluation dataset.

Figure 9 shows the training loss during our finetuning experiment with Whisper medium. Initially, the training loss sharply decreases from approximately 0.4613 at

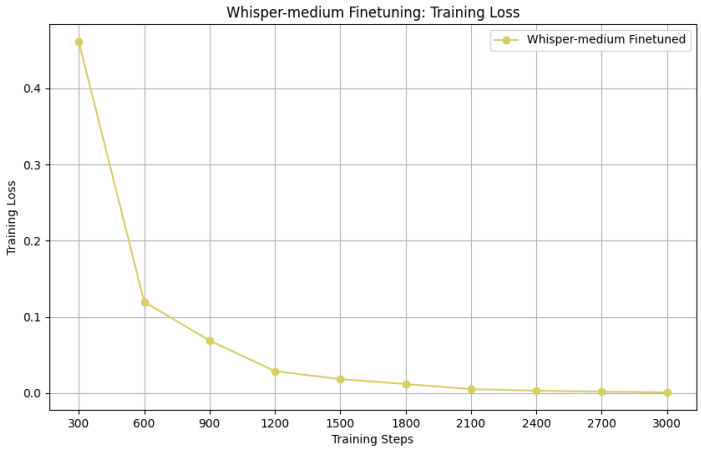

Figure 9: Whisper-medium Finetuning: Training Loss

300 steps to 0.1198 at 600 steps. This rapid reduction suggests that the model was adjusting to the dominant features of the dataset immediately.

As training continued, the loss steadily declined, which indicates ongoing improvements in the model's ability to adapt to the dataset. Notably, the loss values reduced more slowly after the first 900 steps, transitioning from a loss of 0.0689 at 900 steps to a final low of 0.0009 at 3000 steps. This shows how the model transitioning from learning general patterns to refining its understanding of the more subtle, yet crucial, details of the data to achieve high accuracy.

The validation loss during finetuning of Whisper-medium is presented in Figure 10. It reveals an initial loss reduction from 0.1944 to 0.1711 between 300 and 600 steps, which suggests effective early learning. However, the subsequent increase in validation loss from 0.1725 at 900 steps to 0.1921 at 3000 steps suggests potential overfitting as training progresses.

The WER results of the finetuned Whisper-medium model on our testset and the MedASR-DE-test dataset are illustrated in Figure 11 and Figure 12. When using the Whisper-medium model, the WER result before finetuning without normalization for the testset of MedASR-DE-AT was 24.89%, and for the MedASR-DE-test dataset, it was 15.61%. After finetuning without normalization, these rates improved to 10.37% and 9.48%, respectively.

When normalization was applied, the WER results before finetuning were 17.97% for the testset of MedASR-DE-AT and 11.47% for the MedASR-DE-test dataset. After finetuning with normalization, the WER further decreased to 6.77% for the MedASR-DE-AT dataset and 7.19% for the MedASR-DE-test dataset. The results are shown in Figure 13 and Figure 14.

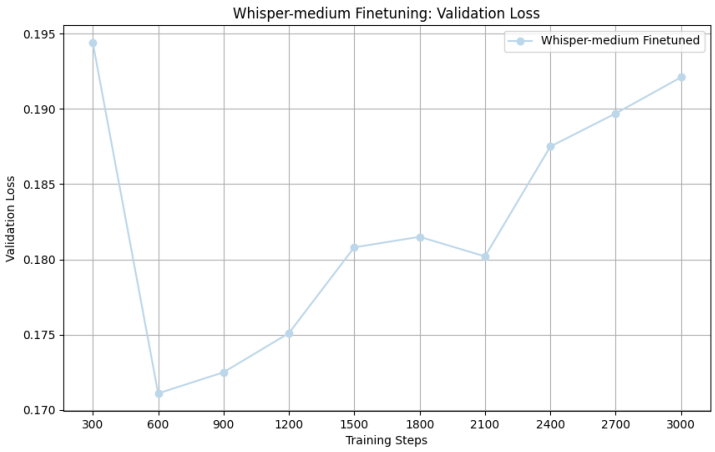

Figure 10: Whisper-medium Finetuning: Validation Loss

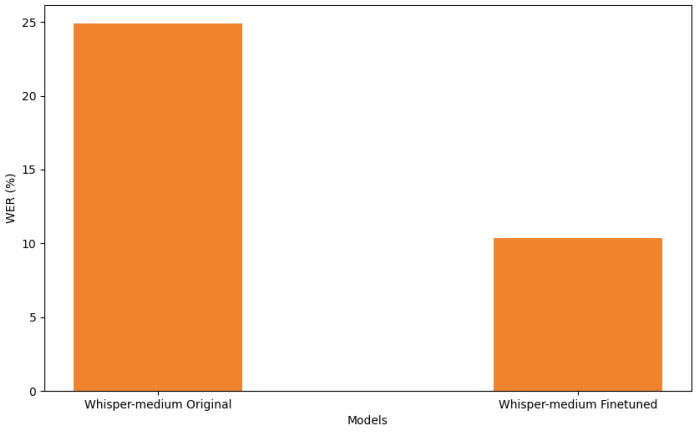

Figure 11: WER (%) ↓ improvements for Finetuned Whisper-medium (without normalization) on test set of MedASR-DE-AT

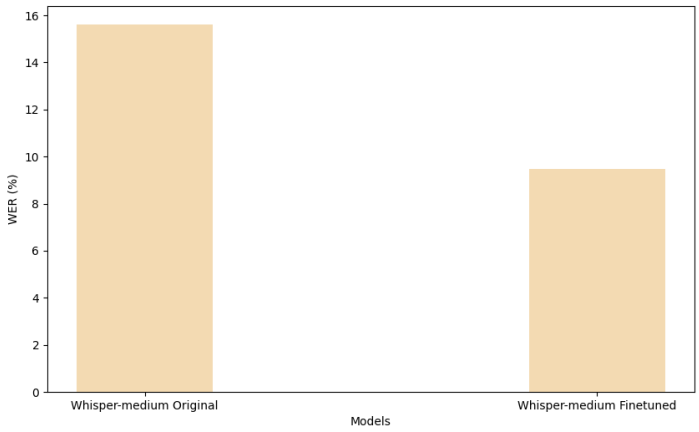

Figure 12: WER (%) ↓ improvements for Finetuned Whisper-medium (without normalization) on MedASR-DE-test

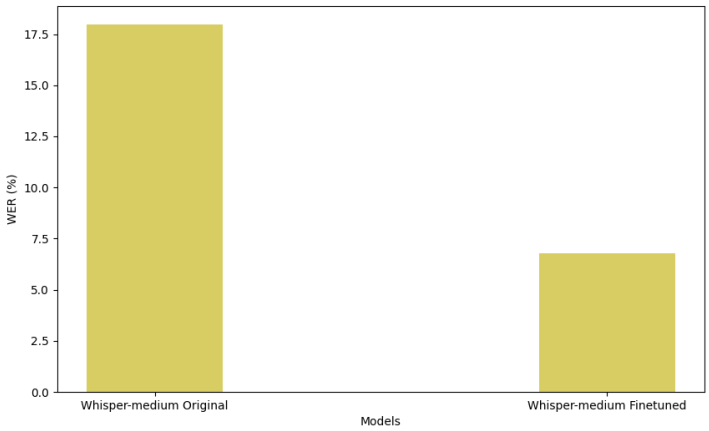

Figure 13: WER (%) ↓ improvements for Finetuned Whisper-medium (with normalization) on test set of MedASR-DE-AT

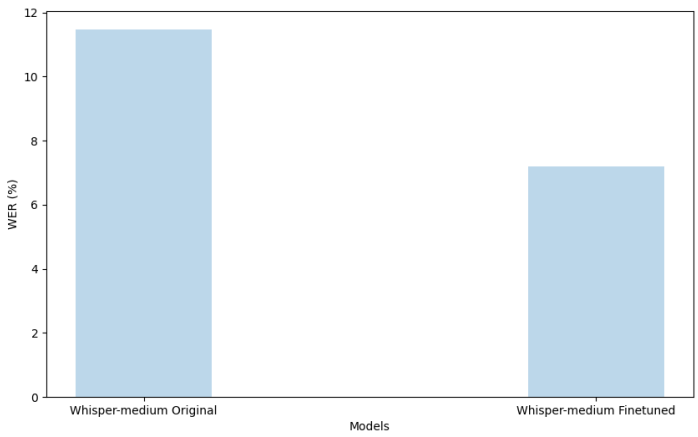

Figure 14: WER (%) ↓ improvements for Finetuned Whisper-medium (with normalization) on MedASR-DE-test

The substantial decrease in training loss and WER during the finetuning process of the Whisper-medium model clearly shows enhancements in the finetuned model's transcription capabilities. Training and finetuning the Whisper-medium model on our German medical dataset with Austrian Standard German pronunciation has effectively enabled the model to become highly adept at processing and transcribing the specific terminology and language patterns represented within the domain of surgery. The consistent reduction in WER, further supported by normalization techniques, confirms the effectiveness of our approach.

Moreover, it is important to highlight that the results on the MedASR-DE-AT dataset, which includes Austrian pronunciation and was used during training, showed more significant improvements. This outcome aligns with expectations given the model's training specifics. Conversely, the improvements on datasets featuring Standard German pronunciation were less pronounced, yet noticeable. This indicates that the finetuning also enhanced the model's ability to transcribe medical terminology more accurately across different dialects of German.

7.2.2 Impact of Training Data Volume on Model Performance

Figure 15 shows the results of our evaluation of the impact of training data volume on the Whisper medium model's performance. The figure illustrates the WER (in percent) against the amount of training data (in hours). Initially, both the model employing normalization and the one without normalization steps exhibit

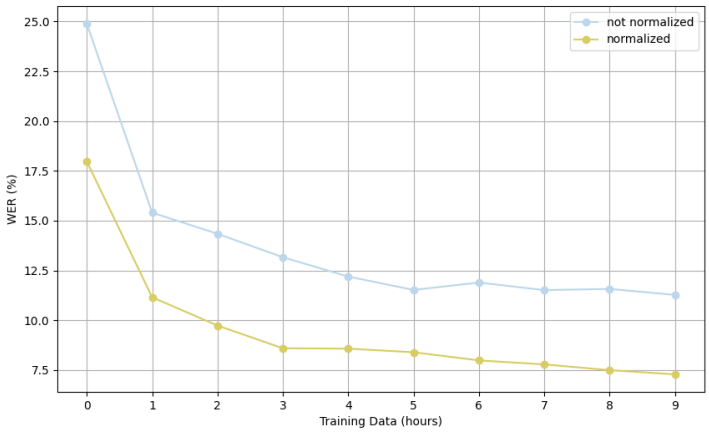

Figure 15: Impact on Training Data Volume on Performance of Whisper-medium on MedASR-DE-AT

a significant drop in WER after the first hour of training. The WER for the model without normalization steps on MedASR-DE-AT decreases from 24.89% to 15.40%, while the model including normalization WER drops from 17.97% to around 11.15%.

With the inclusion of more training data, both models show a gradual decline in WER, with diminishing WER as the amount of training data increases, which shows the model's capacity to adapt to the specific linguistic and acoustic features of the medical domain present in the training data. However, the model including normalization consistently outperforms the model without normalization steps at all training data levels. After 9 hours of training, the WER for MedASR-DE-AT without normalization is 11.27%, whereas the model employing normalization achieves a WER of 7.28%.

The results of the same experiment on the MedASR-DE-test dataset are illustrated in Figure 16. For MedASR-DE-test, after the first hour of training, the WER improves from 15.61% to 13.42% when no normalization is applied, while the model employing normalization initially remains stable around 11.5% WER. As more training data is incorporated, both models exhibit a decline in WER, with the model using normalization once again consistently outperforming the other model at all training data levels. After 9 hours of training, the WER for the model without normalization steps is at 9.48%, whereas the model employing normalization steps achieves a WER of 7.19%. The observed slight upward trend at the conclusion of the training process can be explained by the variability of machine learning models rather than a fundamental issue with the model's performance.

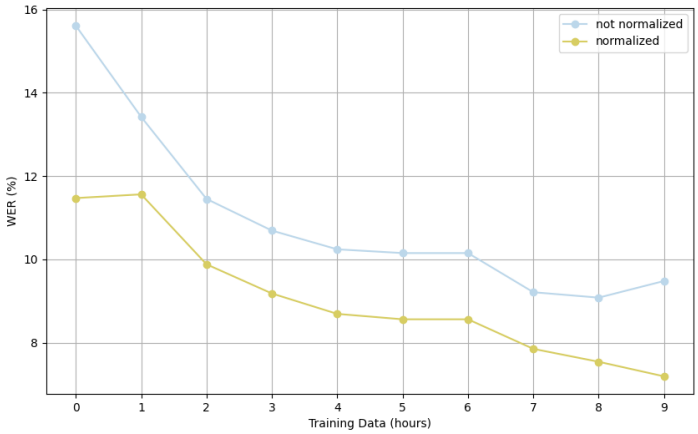

Figure 16: Impact on Training Data Volume on Performance of Whisper-medium on MedASR-DE-test

Given that machine learning models do not produce exactly the same outcomes on separate occasions, this variation can be considered normal. It underscores the importance of repeated testing and validation to accurately assess the model's effectiveness across different datasets and conditions. Such variability is a natural aspect of how machine learning algorithms learn and generalize from data.

Normalization significantly improves model performance, leading to lower WERs at all training data levels and for both datasets. This suggests that the Whisper model benefits from cleaner, more standardized input data. Increasing the amount of training data reduces the WER for both models, although the rate of improvement slows with more data, indicating that additional training data continues to enhance performance but with decreasing incremental gains.

A critical observation is the influence of pronunciation differences between Austrian German and Standard German on the model's performance. The overall WER improvement is once again more pronounced in the MedASR-DE-AT dataset, which includes Austrian German pronunciation, compared to the MedASR-DE-test dataset with Standard German pronunciation, which makes sense, given that we trained the model on the train split of the MedASR-DE-AT dataset, compromising solely Austrian Standard German pronunciation. This suggests that normalization and finetuning are particularly beneficial in handling pronunciation variations, consequently improving the model's robustness and accuracy in diverse linguistic contexts within the medical domain. This observation is further reinforced by the fact that the model was predominantly trained on recordings of the author's voice. Consequently, it demonstrates enhanced accuracy when tested again using similar

vocal characteristics, which further underlines the benefits of personalized training in improving speech recognition performance in specific linguistic and acoustic environments.

In summary, the experiment demonstrates that both increasing the volume of training data and applying normalization techniques substantially improve the Whisper model's accuracy for German medical data transcription, with normalization providing a notable advantage throughout the training process. These results underscore the efficacy of model adaptation through incremental training on domain-specific data. By finetuning the Whisper-medium model with increasing volumes of targeted audio data, we could improve the WER significantly, enhancing the model's performance and reliability for applications within the medical field. The consistent reduction in WER with additional training data also confirms the potential of deep learning models to learn complex patterns in specialized contexts, provided that sufficient and representative training data is used. However, pronunciation differences between training and testing data sets can adversely affect performance, indicating a need for pronunciation-specific training to achieve optimal results in medical speech recognition technologies. This insight provides a valuable foundation for further experiments and applications in medical ASR technologies.

7.2.3 Finetuning Whisper-large-v3

The finetuning of the Whisper-large-v3 model on German audio data demonstrated less notable outcomes than the finetuning experiments of Whisper-medium. This can be attributed to the fact that the runtime on Google Colab was consistently interrupted after approximately 900 steps or approximately four to six hours of training. These interruptions are particularly detrimental to the training of larger models like Whisper-large-v3, which require extended, uninterrupted runtimes to effectively adjust their more complex neural architectures. As a result, the model could not fully optimize its parameters or adequately learn from the training data.

Figure 17 displays the progression of the Whisper-large-v3 model's finetuning until the 900th step, measured by the WER. Initially, the WER starts at 20.31% at 0 training steps. As training continues, a gradual decrease in the WER is observed: it reduces to 20.07% at 300 steps, drops to 19.24% by 600 steps, and further declines to 18.94% at 900 steps.

The training loss during the finetuning experiment of Whisper-large-v3 using AdaLoRA is shown in Figure 18. The results reveal a consistent reduction across training steps. The loss declined from 1.2541 at 300 steps to 1.0263 at 600 steps, and further to 0.6673 at 900 steps. This signifies that the model is effectively learning and improving its capacity to approximate the target outputs, showing an enhanced ability to generalize from the training data. The steady decrease in loss

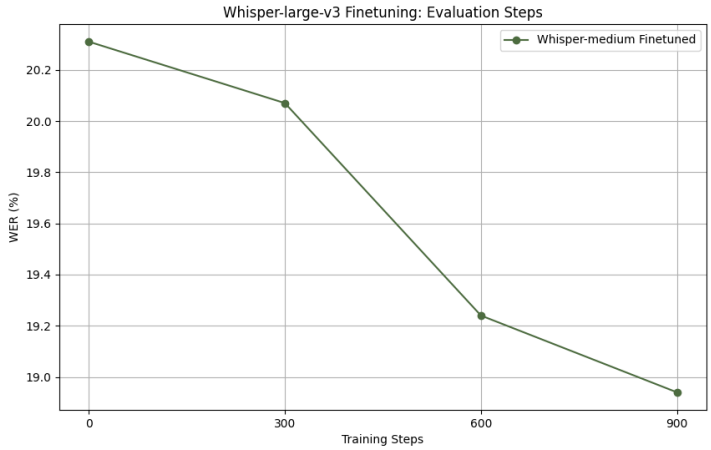

Figure 17: Whisper-large-v3 Finetuning: Evaluation Steps

confirms that the adjustments made through AdaLoRA finetuning are successfully optimizing the model's performance.

The validation loss results for finetuning the Whisper large-v3 model using AdaLoRA show a continuous decrease over the training steps, and are illustrated in 19. Starting at 1.2164 at 300 steps, the loss drops to 0.8764 at 600 steps and further reduces to 0.6093 at 900 steps. This consistent downward trend suggests that the model is successfully generalizing from the training data and enhancing its predictive accuracy on unseen data. The results reflect positively on the AdaLoRA finetuning method. They indicate its efficacy in improving model performance across the validation dataset.

The results of testing the finetuned Whisper-large-v3 model on both the test set of the MedASR-DE-AT dataset and the MedASR-DE-test dataset are presented in Figures 20, 21, 22, and 23. The finetuning results of the Whisper-large-v3 model were also influenced by the application of normalization and the specific datasets used. The results were evaluated by considering changes in the WER before and after finetuning under conditions of normalization and without normalization.

In the normalized condition, the MedASR-DE-AT dataset showed a slight improvement in WER following finetuning, decreasing from 13.62% to 13.33%. Conversely, the MedASR-DE-test dataset, which initially recorded a lower WER of 7.23%, experienced a slight increase to 7.90% after finetuning. This deviation does not necessarily indicate a decline in the model's overall performance through finetuning. Given that the training and validation loss continuously decreased it is more likely that the slight increase in WER, less than 1%, is attributed to the nature of machine learning models and, in this case, their variability in creating

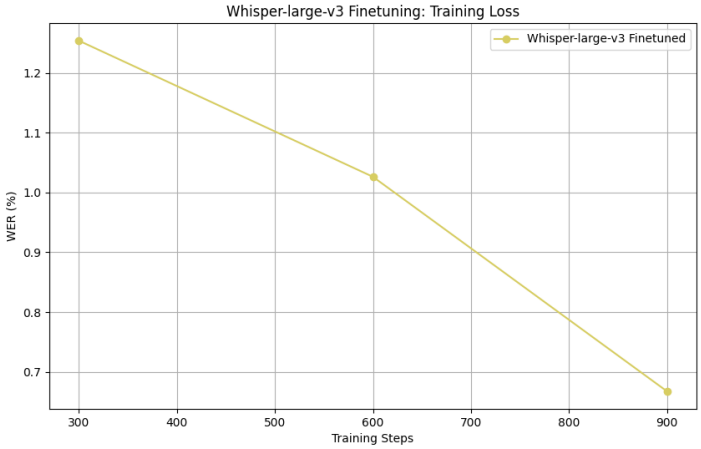

Figure 18: Whisper-large-v3 Finetuned: Training Loss

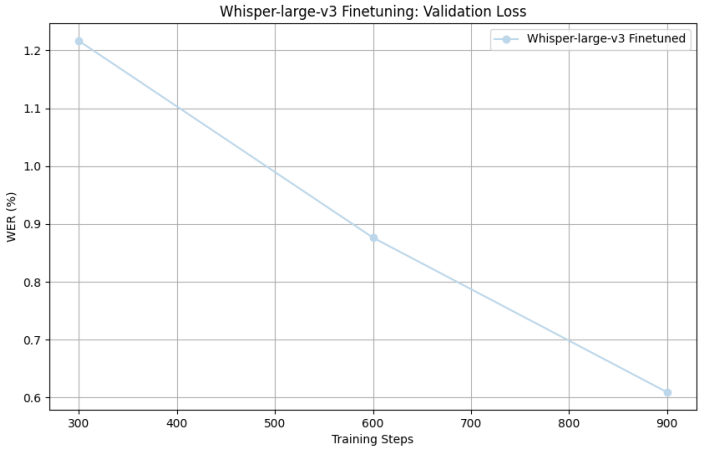

Figure 19: Whisper-large-v3 Finetuned: Validation Loss

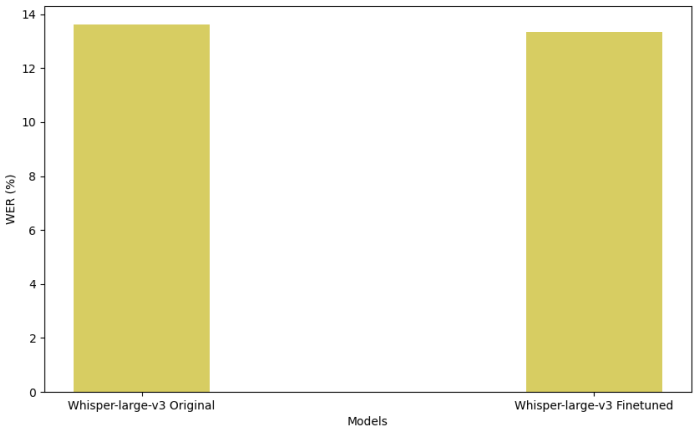

Figure 20: WER (%) ↓ improvements for Finetuned Whisper-large-v3 (with normalization) on test set of MedASR-DE-AT

transcriptions, as described in 7.2.2. The results are illustrated in Figure 20 and Figure 21.

For the non-normalized experiments, the initial WERs were higher, indicating the beneficial effect of normalization in reducing transcription errors. The MedASR-DE-AT dataset showed a WER of 20.88% before finetuning, which was reduced to 19.20% afterwards, demonstrating a benefit from finetuning even without normalization. Similarly, the MedASR-DE-test dataset initially had a WER of 11.67%, which improved to 10.33% post finetuning. The results are shown in Figure 22 and Figure 23.

The finetuning of the Whisper-large-v3 model using AdaLoRA demonstrates that even with the constraints of limited runtime on platforms like Google Colab, the model can still learn and improve significantly. This is evident from the consistent decrease in training and validation losses across the training steps. Normalization clearly helped reduce transcription errors, which is crucial for enhancing accuracy in real-world applications. This suggests that thoughtful preprocessing can play a critical role in improving model outcomes. All the finetuning results can be seen in Table 7. These results underscore the effectiveness of AdaLoRA finetuning in refining large-scale neural network models, but also the practical limitations of current training environments. The full potential of models like Whisper-large-v3 might be better realized with more robust computational resources that permit longer and uninterrupted training sessions.

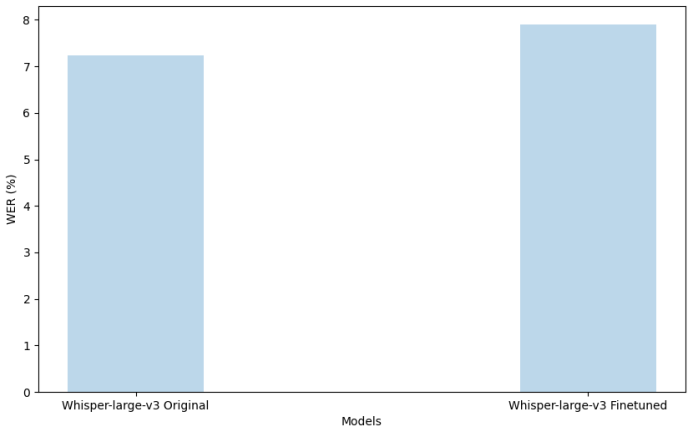

Figure 21: WER (%) ↓ improvements for Finetuned Whisper-large-v3 (with normalization) on MedASR-DE-test

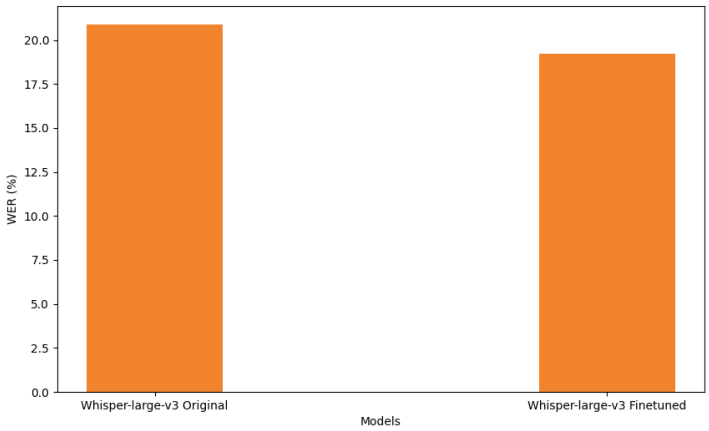

Figure 22: WER (%) ↓ improvements for Finetuned Whisper-large-v3 (without normalization) on test set of MedASR-DE-AT

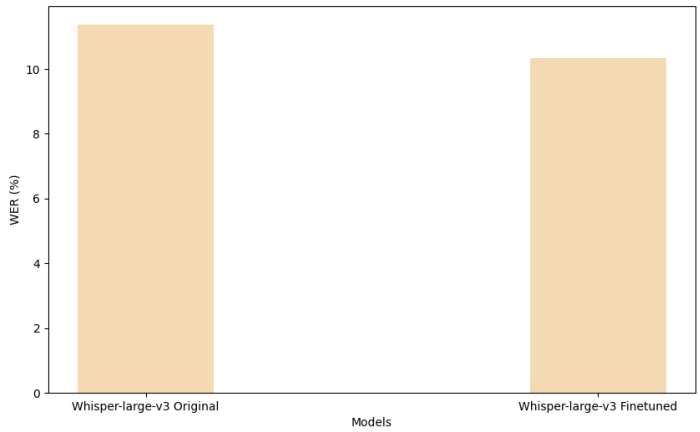

Figure 23: WER (%) ↓ improvements for Finetuned Whisper-large-v3 (without normalization) on MedASR-DE-test

Table 7: WER Results (%) ↓ for fine-tuned Whisper models

Without Normalization		
Description	**Medium**	**Large-V3**
MedASR-DE-AT Original	24.89%	20.21%
MedASR-DE-AT Finetuned	10.37%	19.20%
MedASR-DE-test Original	15.61%	11.67%
MedASR-DE-test Finetuned	9.48%	10.33%
With Normalization		
Description	**Medium**	**Large-V3**
MedASR-DE-AT Original	17.97%	13.62%
MedASR-DE-AT Finetuned	6.77%	13.33%
MedASR-DE-test Original	11.47%	N/A
MedASR-DE-test Finetuned	7.19%	7.90%

7.3 Error Analysis

In this section, we present the findings from our error annotation done with Label Studio.

Original Whisper-medium on MedASR-DE-AT In our analysis of the original Whisper-medium model using the MedASR-DE-AT test set, we found that 213 out of the 158 transcriptions, so approximately 47.54% of them, contained errors. This is a significant error rate which would most probably impact the practical deployment of the model in clinical settings or other sensitive environments, where high-quality medical transcription is needed.

The total word count in the generated transcriptions was 5,825. Out of these, 6.28% of the words were labeled as errors. In total, 366 instances were labelled as errors. Out of that, 264 were error type errors related to medical terms, which accounted for 72.13% of all labels and 4.53% of total words. This indicates a substantial challenge in accurately transcribing medical terminologies, which are crucial for the correct interpretation of medical documentation and communication. This finding aligns with existing challenges in (German) medical ASR, as discussed in Section 1.2.1 and Section 1.3. Other error categories such as insertions, omissions, notation errors, misrecognitions, orthographic errors, grammatical errors, numerical errors, and errors involving abbreviations were considerably less frequent. Specifically, notation errors and misrecognitions comprised 9.29% and 7.38% of the total labels respectively. The label "ORHTH", for orthographic errors appears 20 times, constituting only 0.34% of all words in the dataset and 5.46% of the total labels. Following this, grammatical errors are recorded 15 times, making up 0.26% of total words and accounting for 4.10% of all labels. The omission label occurs 4 times, which represents 0.07% of the total word count and 1.09% of the labels. Both insertions and numerical errors are noted only once, each comprising 0.02% of the words and 0.27% of the labels. Finally, the label "ABBR" has no occurrences, indicating that no abbreviation has been transcribed wrongly. Figure 24 shows the most common error types identified during the analysis of transcriptions performed using the Whisper-medium model on the MedASR-DE-AT dataset.

Figure 25 presents the five most common transcription errors labeled as Medical Terms in the original Whisper-medium model's analysis using the MedASR-DE-AT dataset. These errors include misrecognized words such as *Druck H* mislabeled as *Trokar*, *Overhold* instead of *Overholt*, *Druckhase* for *Trokar*, *Lure* as *Luer*, and *Trocars* labeled as *Tokars*.

These errors can be explained by the characteristics of Austrian pronunciation, as discussed in Section 1.3.2. For example, the confusion between *Druck H* and *Trokar* involves the misrecognition of initial consonants *d* and *t*, a typical issue in Austrian Standard German where such consonants can be variably articulated.

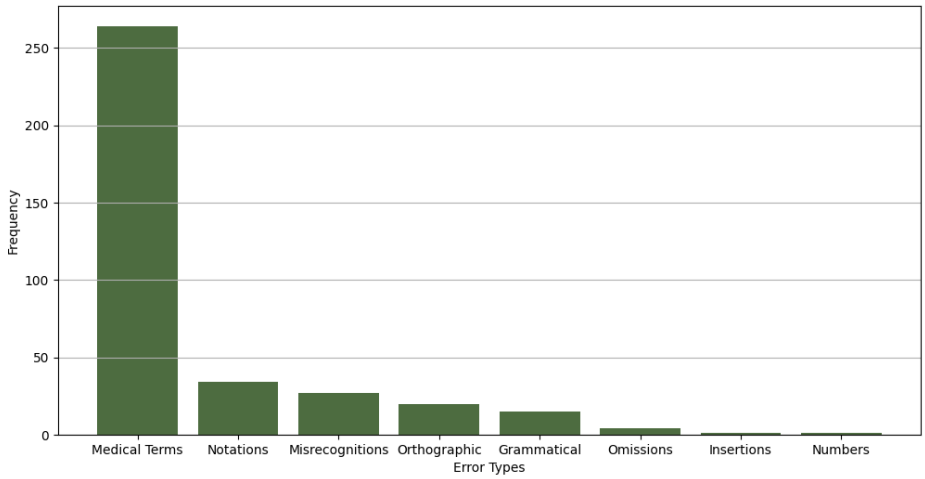

Figure 24: Most Common Labels: Original Whisper-medium (MedASR-DE-AT)

Similarly, the replacement of *Overhold* with *Overholt* and the ending *-er* in *Luer* getting dropped to become *Lure* align with common Austrian phonetic tendencies where final consonants might be softened or omitted. These findings affirm the expected challenges in accurately transcribing such terms without tailored model training to account for these regional variations.

Overall, the transcription errors depicted align with anticipated pronunciation challenges inherent to Austrian German, reinforcing the need for targeted improvements in the model's training regimen to enhance its accuracy in medical settings specific to this linguistic context.

The distribution of errors underscores the need for targeted refinements in the model, particularly to improve the recognition of medical terms, where error rates are notably higher. This analysis suggests that although the Whisper-medium model has potential, it demands specific enhancements to reliably function in German medical contexts.

Finetuned Whisper-medium on MedASR-DE-AT After finetuning the Whisper Medium model with our specialized German medical dataset, we observed a substantial reduction in the error rate among the transcriptions. The percentage of transcriptions containing errors decreased significantly from 47.54% to 20.63%. This improvement highlights the effectiveness of our approach to enhance the accuracy of ASR systems in specialized fields such as medicine.

The total number of labels also declined significantly to 2.02% of the total words, compared to 6.28% in the original Whisper model, which indicates a more reliable

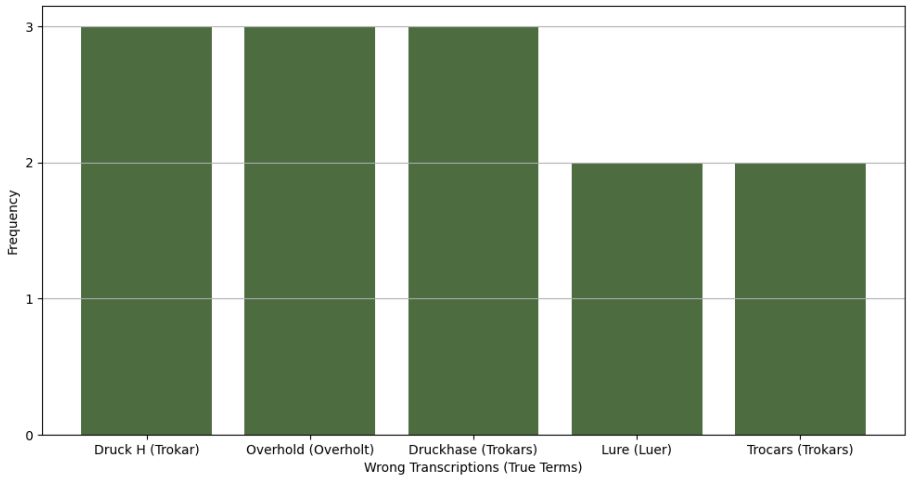

Figure 25: Most Common MEDTERM Errors: Original Whisper-medium (MedASR-DE-AT)

transcription performance across various types of errors. The analysis of labels after finetuning reveals a significant reduction in errors related to Medical Terms, which now makes up 0.89% of total words and 44.04% of total labels. This is a marked improvement from 4.53% of total words and 72.13% of all labels in the original model. Notably, we observed that all of the terms depicted in Figure 25 that were incorrectly transcribed by the original model, now were correctly reconized by our new, finetuned model, which suggests a significant enhancement of accuracy in transcribing medical terminology. After medical terms, errors involving notation appeared most frequently with 21 occurrences, making up 0.39% of the total words and 19.27% of the total labels. This is followed by grammatical errors, which are recorded 17 times, accounting for 0.32% of total words and 15.60% of total labels. Misrecognitions are also notable, with 10 occurrences, constituting 0.19% of the total words and 9.17% of total labels. orthographic errors are observed 9 times, representing 0.17% of total words and 8.26% of total labels. Insertions and omissions show much lower frequencies, with insertions occurring only 2 times, representing 0.04% of total words and 1.83% of total labels, and omissions only once, representing 0.02% of total words and 0.92% of total labels. Only one numerical error appeared, amounting to 0.02% of total words and 0.92% of total labels. Abbreviation errors did not occur at all. This can be seen in Figure 26.

We observed that no individual Medical Term error occurred more than once as an error. This indicates a considerable improvement in the model's ability to accurately recognize and transcribe medical terminology. Consequently, creating

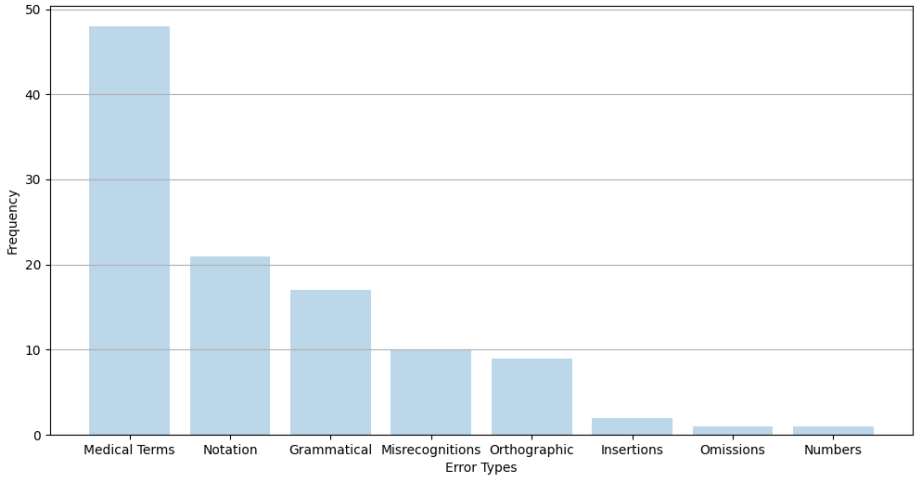

Figure 26: Most Common Labels: Original Whisper-medium (MedASR-DE-AT)

a graph to depict the frequency of individual medical term errors would not be meaningful, as each term appears only once.

These results show significant improvements in accuracy, especially in the critical area of medical terms. This supports the need for ongoing efforts in dataset enhancement and model tuning to further optimize ASR utility in German-speaking healthcare sectors.

Original Whisper-medium on MedASR-DE-test In the analysis of the original Whisper Medium model using the MedASR-DE-test dataset with German pronunciation, 75 out of the 158 total transcriptions, so 47.47% contained errors, which closely aligns with the error rate found using an Austrian pronunciation dataset.

The total word count in the generated transcriptions was 2,259, and 103 words, so 4.56% of the words, were labeled with errors. This percentage represents a slight improvement over the results with the Austrian dataset, indicating that the original the model handles standard German pronunciation sightly better then Austran pronunciation.

MEDTERM errors were again the most frequent, accounting for 3.50% of total words and 76.70% of all labeled errors, which aligns with the dominant error type in the Austrian dataset and underscores the model's persistent difficulty with medical terminology.

The second most common error category is grammatical errors, with 10 occurrences, making up 0.44% of total words and 9.71% of total labels. This is followed

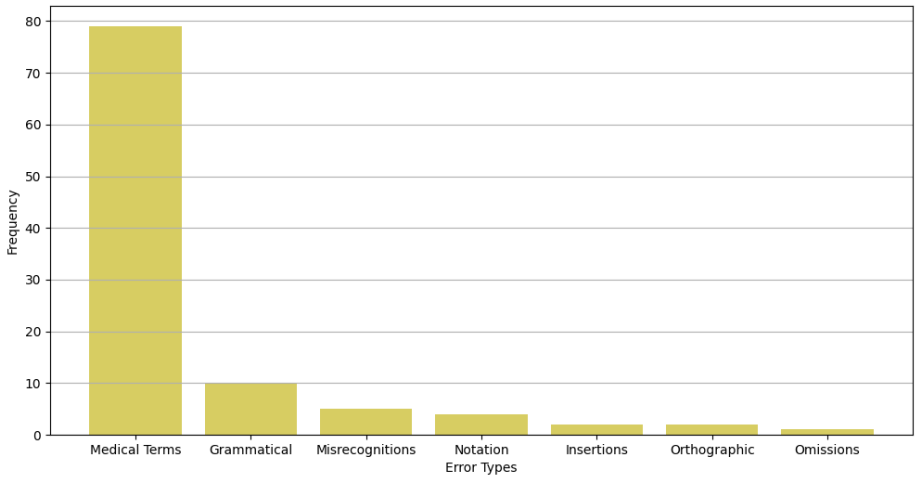

Figure 27: Most Common Labels: Original Whisper-medium (MedASR-DE-test

by misrecognitions, which are noted 5 times and account for 0.22% of total words and 4.85% of total labels. Errors in notation are recorded 4 times, comprising 0.18% of total words and 3.88% of total labels. Errors categorized as insertions and orthographic errors are both observed 2 times, each constituting 0.09% of total words and 1.94% of total labels, indicating minor issues in terms of additional unintended elements and spelling accuracy. Omissions, where elements are missing from the transcription, occur only once, representing 0.04% of total words and 0.97% of total labels. Notably, numerical errors and errors with abbreviations do not appear in the dataset. Figure 27 shows the distribution of labeled errors.

Figure 28 displays the most common transcription errors categorized as medical terms, identified when using the original Whisper-medium model on the MedASR-DE-test dataset, which features German pronunciation.

The errors highlighted include incorrect transcriptions of medical terms, with the correct terms in parentheses for reference: *Vikrül* (*Vicryl*), *Traktos* (*Tractus*), *Vigryl* (*Vicryl*), *bongiosa* (*spongioser*), and *Strecksäne* (*Strecksehne*).

This analysis underscores the specific pronunciation-related challenges that the Whisper-medium model encounters with German medical terms.

Enhancing the model's sensitivity to specialized vocabulary and complex linguistic structures remains crucial for its reliable application in medical fields.

Finetuned Whisper-medium on MedASR-DE-test When testing our finetuned model on the Standard German dataset, we observed that 64 transcriptions, so 40.51% across the total 158 transcriptions, contained errors. This error rate

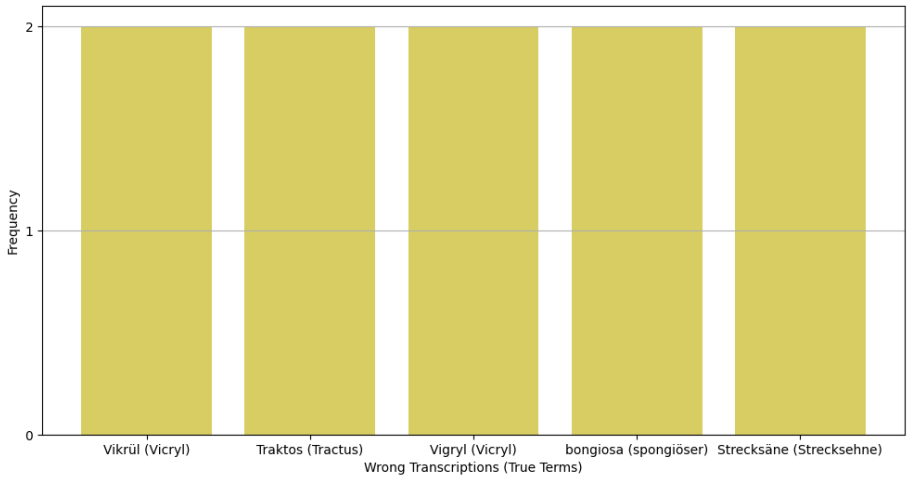

Figure 28: Most Common MEDTERM Errors: Original Whisper-medium (MedASR-DE-test

represents a decrease of 7% compared to the original model's performance on the same German dataset, where the error rate was 47.47%. This suggests that finetuning with the Austrian dataset has imparted some improvements in the model's overall capabilities, and especially in correctly recognizing medical terms, despite the dialectal differences between the training and testing datasets. Here, labeled errors accounted for 3.66% of all words. This is a reduction from 4.56% in the pre-finetuned model tested on German, which indicates enhanced accuracy after the finetuning experiment.

Our analysis shows that MEDTERM errors continue to be the most common, comprising 1.79% of total words and 48.78% of all labeled errors. This proportion highlights a persistent challenge in accurately transcribing medical terminology, though the frequency of these errors has decreased from 3.50% and 44.04%, respectively, in the original model.

After that, the two most frequent errors identified are misrecognitions and grammatical errors, each recorded 11 times. These errors represent 0.49% of total words and account for 13.41% of total labels. Orthographic errors and errors involving notation both appeared 8 times, making up 0.36% of total words and 9.76% of total labels. Less frequent errors include omissions and insertions, each occurring 2 times and representing only 0.09% of total words and 2.44% of total labels. There are no errors related to numbers or abbreviations. The numbers are illustrated in Figure 29

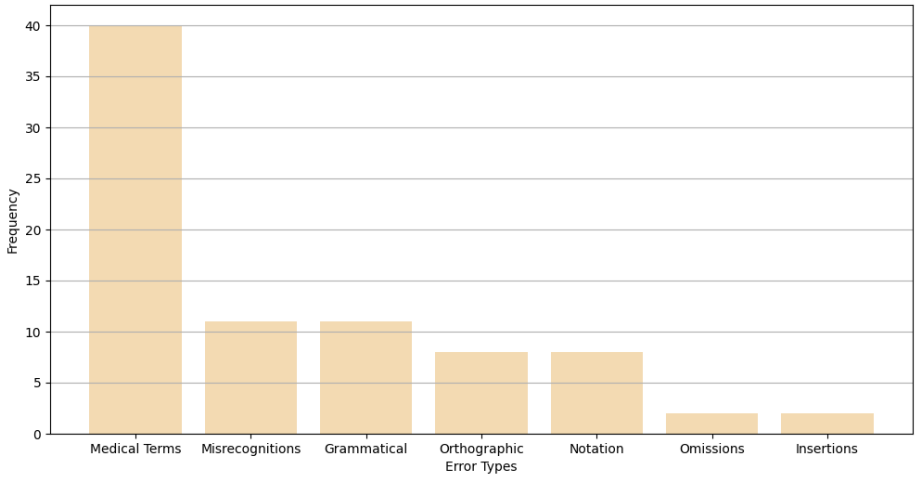

Figure 29: Most Common Labels: Finetuned Whisper-medium (MedASR-DE-test)

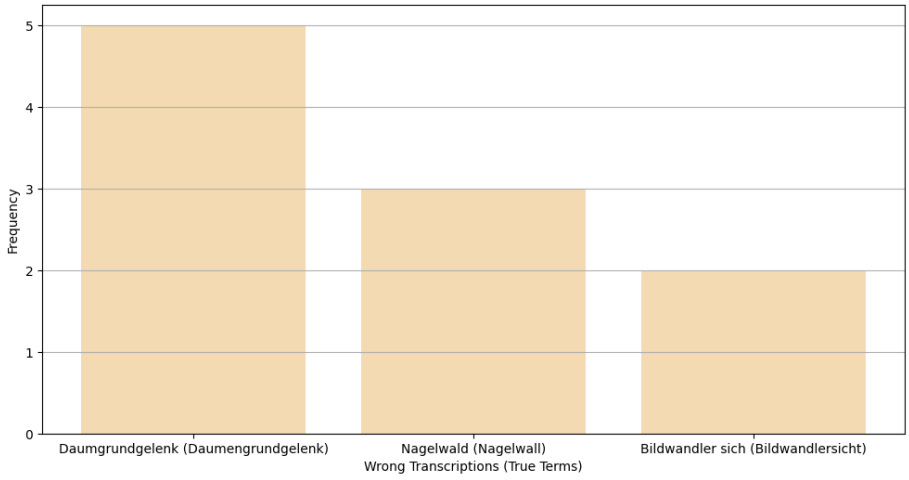

Figure 30: Most Common MEDTERM Errors: Finetuned Whisper-medium (MedASR-DE-test)

Figure 30 shows the frequency of the most common transcription errors categorized as medical terms when using the original Whisper-medium model on the MedASR-DE-test dataset. The term *Daumengrundgelenk* is misrecognized as *Daumgrundgelenk* with the highest frequency among the errors, appearing five times. *Nagelwald*, an incorrect transcription for *Nagelwall*, appears three times. Similarly, *Bildwandler sich* is incorrectly transcribed for *Bildwandlersicht*, appearing twice. This specific error might be due to the complexity of the compound term and the acoustic similarity between the terms in German.

These findings illustrate that while finetuning the Whisper model with a dataset characterized by a different dialect can improve performance, challenges remain, particularly in handling complex medical terminology. The reduced error rates in various categories, however, confirm the benefit of finetuning.

8 Discussion

Similar to (Pekarek-Rosin and Wermter, 2023) and (Jain et al., 2023), we employ finetuning of pre-trained models on domain-specific data to enhance ASR performance, underscoring the effectiveness of specialized finetuning in addressing domain-specific challenges. Like (Wirth and Peinl, 2023; Rumberg et al., 2022; Ardila et al., 2019), we propose a new German dataset. However, we introduce distinct contributions to the field of ASR domain adaptation. Firstly, the focus lies on the German medical domain, more specifically on Austrian Standard German pronunciation within the domain of surgery. Since this is a rather unexplored area within the context of ASR we created a novel dataset specifically tailored for this purpose. This dataset represents a unique contribution by addressing the underrepresented area of medical ASR in the (Austrian Standard) German language (Wirth and Peinl, 2022).

While Pekarek-Rosin and Wermter (2023) explore layer-specific finetuning and Experience Replay strategies, we focus on adapting finetuning hyperparameters without continual learning techniques. However, we explore the use of PEFT (Houlsby et al., 2019b) and AdaLoRA (Zhang et al., 2023) for enhancing efficiency in our approach. For evaluation, we use the same metric as both of the research teams, the WER, which is most commonly used for evaluating ASR tasks (von Neumann et al., 2022).

Additionally, we incorporate a detailed evaluation of the Whisper model's performance before and after domain adaptation, offering insights into the efficacy of finetuning strategies for medical terminology recognition. Furthermore, we offer detailed understandings of model performance improvements and the specific types of errors addressed through domain adaptation by performing error annotation and analysis through Label Studio.

9 Limitations

While this thesis makes notable contributions to the field of ASR in the German medical contexts, it also includes several limitations that could potentially influence the outcomes and generalizability of the findings.

One of the primary limitations was the requirement for a substantial amount of audio data for effective domain adaptation. The creation of the surgery-specific German medical dataset was particularly challenging due to the need to manually record a large number of the audio files. This process was time-consuming and intensive, since it only involved two individuals. Out of the two speakers, the author of this thesis recorded a vast majority, which could potentially affect the diversity and representativeness of the dataset. Furthermore, the lack of a medical background among the people who recorded the audios for the dataset necessitated additional steps to verify the pronunciation and meaning of medical terms, which introduced another layer of complexity and potential for error to the thesis.

Another significant limitation was related to the computational resources available for training and finetuning the ASR models. We had to rely on Colab Pro for computational power, which presented challenges in terms of limited runtime and the costs associated with GPU usage. The restriction on continuous runtime meant that experiments could not be conducted over extended periods, which is typically necessary for training robust deep learning models. Moreover, the financial implications of acquiring additional compute hours restricted the scope of experimental trials that could be performed.

10 Conclusions

In this thesis, we have demonstrated that domain-specific finetuning can substantially improve model performance on specialized tasks, such as German medical ASR. Notably, by creating and utilizing our own dataset tailored to the nuances of Austrian Standard German, we were able to reduce the WER of the Whisper-medium model on the MedASR-DE-AT dataset by 14.52%.

Furthermore, we have shown that finetuning the Whisper-medium model with a dataset featuring Austrian Standard German pronunciation significantly improves its transcription accuracy. The percentage of transcriptions containing errors in the Austrian dataset declined importantly by 26.91%, which presents a substantial enhancement. Moreover, finetuning with Austrian speech data also positively impacted the model's performance on the Standard German dataset, where the error rate decreased by 6.13%. While this improvement is not as pronounced as it was for the Austrian dataset, it is still noteworthy. These findings suggest that ASR systems can in fact greatly benefit from targeted training on specific dialects.

In conclusion, significant progress has been made in adapting ASR systems to the German-speaking medical field. This thesis adds to the existing knowledge on ASR technology and lays the groundwork for future research aimed at developing more robust and precise ASR systems for healthcare and other essential sectors, particularly for less-studied language variants and dialects.

10.1 Future Work

In this chapter, we outline the steps and improvements that could be made in future work based on the findings of this thesis.

In Section 7.2.1, we mention that the validation loss increased slightly when finetuning Whisper-medium for 3 epochs. Despite the increasing validation loss, the consistent decline in both training loss and WER shows that the model is improving on the training data. However, this improvement does not fully extend to the validation data, indicating that while the model is becoming more adept at handling the training dataset's specific characteristics, it struggles to maintain this performance on unseen data, typical of overfitting.

To address this issue and enhance the model's ability to generalize, several strategies could be considered. Implementing early stopping could help preserve the

model at its optimal state of generalization by halting training when the validation loss begins to rise. Enhancing regularization, through methods such as dropout or L2 regularization, could discourage the model from fitting too closely to the noise in the training data. Adjusting the learning rate, either by reducing it as training progresses or employing a learning rate schedule that decreases the rate over time, can ensure finer adjustments in the model's parameters during later training phases. Additionally, introducing variations through data augmentation could help the model generalize better to new, unseen examples. Employing cross-validation techniques would also provide a better assessment of the model's performance across different subsets of the data, informing more robust training strategies.

To achieve more substantial improvements and realize the full potential of finetuning the Whisper-large-v3 model, as discussed in Section 7.2.3, a more robust GPU setup is essential. Such a setup would facilitate longer training sessions without disruptions, allowing the model to better adapt to the nuances of the German language audio data, which enhances its accuracy and performance in real-world applications. Securing more robust and dedicated computational resources would enable longer training sessions and more extensive experimentation. A potential solution could be to explore alternative funding sources or partnerships which could alleviate financial constraints and support the continuous development and refinement of ASR systems tailored for medical use.

Additionally, future work in this area should expand existing ASR datasets to include a broader range of speakers and medical scenarios to enhance the diversity and representativeness of the training material. This expansion would likely improve the model's ability to generalize across different accents, dialects, and medical subfields. Researchers should consider collaborations with medical institutions to ensure the accuracy and relevance of the medical terms and their pronunciation in order to create valuable medical ASR datasets.

Acronyms

AdaLoRA Adaptive Low-Rank Adaptation. 27, 53, 54, 66, 67, 69, 81

ASG Austrian Standard German. 8, 40

ASR Automatic Speech Recognition. 1–12, 16–18, 21–24, 29–33, 36, 41–43, 45, 46, 49, 57, 66, 72, 73, 75, 81, 83, 85, 86

CNN Convolutional Neural Network. 2

CV Computer Vision. 15

DA Domain Adaptation. 15

DNN Deep Neural Network. 2

EHR electronic health records. 1, 5

LoRA Low-Rank Adaptation. 27, 28

MFCCs Mel-Frequency Cepstral Coefficients. 2

NeurIPS Neural Information Processing Systems (after rebranding). 14

NIPS Neural Information Processing Systems. 14

NLP Natural Language Processing. 3, 15–22, 24, 50

PEFT Paramater-Efficient Finetuning. 27, 28, 81

RNN Recurrent Neural Network. 2

SOTA State-of-the-art. 5, 41, 43

TL Transfer Learning. 13, 14

WER Word Error Rate. 2, 5, 7, 11, 22–24, 30, 32, 41, 49, 50, 57–60, 63–67, 69, 81, 85

Bibliography

Neurips 2024. Code of Conduct. https://nips.cc/Conferences/2023/CodeOfConduct, note = Accessed: 2024-04-08.

Distil-Whisper. https://github.com/huggingface/distil-whisper.

Rajesh Kumar Aggarwal and Mayank Dave (2011). Acoustic modeling problem for automatic speech recognition system: conventional methods (part I). *Int. J. Speech Technol.*, 14(4):297.

Purvi Agrawal and Sriram Ganapathy (2019). Modulation filter learning using deep variational networks for robust speech recognition. *IEEE J. Sel. Top. Signal Process.*, 13(2):244–253.

Ahmed Ali and Steve Renals (2020). Word error rate estimation without ASR output: e-wer2. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 616–620. ISCA.

Massih-Reza Amini, Vasilii Feofanov, Loïc Pauletto, Emilie Devijver, and Yury Maximov (2022). Self-training: A survey. *CoRR*, abs/2202.12040.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber (2019). Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber (2020a). Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber (2020b). Common voice: A massively-multilingual speech corpus. In

Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pages 4218–4222. European Language Resources Association.

John P Avendano, Daniel O Gallagher, Joseph D Hawes, Joseph Boyle, Laurie Glasser, Jomar Aryee, Brian M Katt, Joseph Hawes, Joseph J Boyle, and Jomar N Aryee (2022). Interfacing with the electronic health record (ehr): a comparative review of modes of documentation. *Cureus*, 14(6).

Uğur Ayvaz, Hüseyin Gürüler, Faheem Khan, Naveed Ahmed, Taegkeun Whangbo, and Abdusalomov Bobomirzaevich (2022). Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. *CMC-computers materials & continua*, 71(3).

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli (2022). XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Samar Bashath, Nadeesha Perera, Shailesh Tripathi, Kalifa Manjang, Matthias Dehmer, and Frank Emmert-Streib (2022). A data-centric review of deep transfer learning with applications to text data. *Inf. Sci.*, 585:498–528.

Suzanne V Blackley, Jessica Huynh, Liqin Wang, Zfania Korach, and Li Zhou (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the american medical informatics association*, 26(4):324–338.

Suzanne V Blackley, Valerie D Schubert, Foster R Goss, Wasim Al Assad, Pamela M Garabedian, and Li Zhou (2020). Physician use of speech recognition versus typing in clinical documentation: a controlled observational study. *International Journal of Medical Informatics*, 141:104178.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Pau Panareda Busto and Juergen Gall (2017). Open set domain adaptation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 754–763. IEEE Computer Society.

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I. Jordan (2018). Partial transfer learning with selective adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2724–2732. Computer Vision Foundation / IEEE Computer Society.

Jungwon Chang, Hosung Nam, et al. (2023). Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on whisper model and ksponspeech dataset. *Phonetics and Speech Sciences*, 15(3):83–88.

Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong (2018). Progressive joint modeling in unsupervised single-channel overlapped speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(1):184–196.

Moonsun Choi (2023). A corpus-based comparative analysis of english speeches used in interpreter training programs in korea and china. In *Corpora in Interpreting Studies*, pages 251–269. Routledge.

Chenhui Chu and Rui Wang (2018). A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1304–1319. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2426–2430. ISCA.

Hal Daumé (2007). Frustratingly easy domain adaptation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Jorge Dávila-Chacón, Jindong Liu, and Stefan Wermter (2019). Enhanced robot speech recognition using biomimetic binaural sound source localization. *IEEE Trans. Neural Networks Learn. Syst.*, 30(1):138–150.

Marc Deisenroth, Aldo Faisal, and Cheng Ong (2020). *Mathematics for Machine Learning*.

Keqi Deng, Zehui Yang, Shinji Watanabe, Yosuke Higuchi, Gaofeng Cheng, and Pengyuan Zhang (2022). Improving non-autoregressive end-to-end speech recognition with pre-trained acoustic and language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8522–8526. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zarnigor Djalilova et al. (2023). The use of latin terminology in medical case. , 2(14):9–15.

Katja Dobrić (2013). Creating medical terminology: from latin and greek influence to the influence of english as the current lingua franca of medical communication. *JAHR–European Journal of Bioethics*, 4(1):493–502.

Onno Eberhard and Torsten Zesch (2021). Effects of layer freezing when transferring deepspeech to new languages. *CoRR*, abs/2102.04097.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. (2020). Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane (2015a). Automatic speech recognition errors detection and correction: A review. In *1st International Conference on Natural Language and Speech Processing, ICNLSP 2015, Algiers,*

Algeria, October 18-19, 2015, volume 128 of *Procedia Computer Science*, pages 32–37. Elsevier.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane (2015b). Automatic speech recognition errors detection and correction: A review. In *1st International Conference on Natural Language and Speech Processing, ICNLSP 2015, Algiers, Algeria, October 18-19, 2015*, volume 128 of *Procedia Computer Science*, pages 32–37. Elsevier.

Maxine Eskenazi, Jack Mostow, and David Graff (1997). The cmu kids corpus. *Linguistic Data Consortium*, 11.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia (2021). A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Antonio Javier Gallego, Jorge Calvo-Zaragoza, and Robert B. Fisher (2021). Incremental unsupervised domain-adversarial training of neural networks. *IEEE Trans. Neural Networks Learn. Syst.*, 32(11):4864–4878.

Sriram Ganapathy (2017). Multivariate autoregressive spectrogram modeling for noisy speech recognition. *IEEE Signal Process. Lett.*, 24(9):1373–1377.

Les Gasser, Kiran Lakkaraju, Sylvian Ray, and Samarth Swarup (2005). Darpa baa 05-29: Transfer learning issues and potential contributions.

Gábor Gosztolya and Tamás Grósz (2016). Domain adaptation of deep neural networks for automatic speech recognition via wireless sensors. *Journal of Electrical Engineering*, 67(2):124–130.

Calbert Graham and Nathan Roll (2024). Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

Michael Gref, Nike Matthiesen, Christoph Schmidt, Sven Behnke, and Joachim Köhler (2022). Human and automatic speech recognition performance on german oral history interviews. *CoRR*, abs/2201.06841.

Hanwen Guan, Haoyi Song, Yukun Yang, and Jiayuan Zhang (2021). The fourier transform and its application in solving the equation. In *Journal of Physics: Conference Series*, volume 2012, page 012056. IOP Publishing.

Kaylynn Gunter, Charlotte Vaughn, and Tyler Kendall (2021). Contextualizing/s/retraction: Sibilant variation and change in washington dc african american language. *Language Variation and Change*, 33(3):331–357.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris (2020). A broader study of cross-domain few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 124–141. Springer.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort (2023). Continual pre-training of large language models: How to (re)warm your model? *CoRR*, abs/2308.04014.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *CoRR*, abs/2403.14608.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève (2018). TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno (2015). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 23(2):373–382.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019a). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019b). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long*

Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli (2021). Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 721–725. ISCA.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zimeng Huang, Bo Jiang, Tian Guo, and Yunzhuo Liu (2023). Measuring the impact of gradient accumulation on cloud-based distributed training. In *23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2023, Bangalore, India, May 1-4, 2023*, pages 344–354. IEEE.

Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He (2022). Large-scale ASR domain adaptation using self- and semi-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6627–6631. IEEE.

Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu (2023). Adaptation of whisper models to child speech recognition. *CoRR*, abs/2307.13008.

Dalwon Jang, Jaewon Lee, and Jong-Seol Lee (2021). Acoustic feedback detection for online video conferencing. In *International Conference on Information and Communication Technology Convergence, ICTC 2021, Jeju Island, Korea, Republic of, October 20-22, 2021*, pages 1516–1518. IEEE.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim (2022). Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Hongpeng Jin, Wenqi Wei, Xuyu Wang, Wenbin Zhang, and Yanzhao Wu (2023). Rethinking learning rate tuning in the era of large language models. In *5th IEEE International Conference on Cognitive Machine Intelligence, CogMI 2023, Atlanta, GA, USA, November 1-4, 2023*, pages 112–121. IEEE.

Haotian Ju, Dongyue Li, and Hongyang R. Zhang (2022). Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10431–10461. PMLR.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos (2021). UDALM: unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2579–2590. Association for Computational Linguistics.

Wouter M Kouw and Marco Loog (2018). An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

Manoj Kumar, So Hyun Kim, Catherine Lord, Thomas D. Lyon, and Shrikanth Narayanan (2020). Leveraging linguistic context in dyadic interactions to improve automatic speech recognition for children. *Comput. Speech Lang.*, 63:101101.

Yogesh Kumar (2024). A comprehensive analysis of speech recognition systems in healthcare: Current research challenges and future prospects. *SN Comput. Sci.*, 5(1):137.

Ngoc Lai (2022). LMN at SemEval-2022 task 11: A transformer-based system for English named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1438–1443, Seattle, United States. Association for Computational Linguistics.

Ho Yong Lee, Ji-Won Cho, Minook Kim, and Hyung-Min Park (2016). Dnn-based feature enhancement using doa-constrained ICA for robust speech recognition. *IEEE Signal Process. Lett.*, 23(8):1091–1095.

Minhyeok Lee (2023). GELU activation function in deep learning: A comprehensive mathematical analysis and performance. *CoRR*, abs/2305.12073.

Sheng-Chieh Lee, Jhing-Fa Wang, and Miao-Hia Chen (2018). Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions. *Sensors*, 18(7):2068.

Tso-Ying Lee, Chin-Ching Li, Kuei-Ru Chou, Min-Huey Chung, Shu-Tai Hsiao, Shu-Liu Guo, Lung-Yun Hung, and Hao-Ting Wu (2023). Machine learning-based speech recognition system for nursing documentation – a pilot study. *International Journal of Medical Informatics*, 178:105213.

Brian Lester, Rami Al-Rfou, and Noah Constant (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Siting Liang, Mareike Hartmann, and Daniel Sonntag (2023). Cross-domain german medical named entity recognition using a pre-trained language model and unified medical semantic types. In *Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023*, pages 259–271. Association for Computational Linguistics.

Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da-Shan Shiu (2023). Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Julian Linke, Saskia Wepner, Gernot Kubin, and Barbara Schuppler (2023). Using kaldi for automatic speech recognition of conversational austrian german. *CoRR*, abs/2301.06475.

Xugang Lu, Sheng Li, and Masakiyo Fujimoto (2020). Automatic speech recognition. In Yutaka Kidawara, Eiichiro Sumita, and Hisashi Kawai, editors, *Speech-to-Speech Translation*, Springer Briefs in Computer Science, pages 21–38. Springer.

Hao Ma, Zhiyuan Peng, Mingjie Shao, Jing Li, and Ju Liu (2023). Extending whisper with prompt tuning to target-speaker ASR. *CoRR*, abs/2312.08079.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom (2021). Automatic speech recognition: a survey. *Multim. Tools Appl.*, 80(6):9411–9457.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria (2023). A review of deep learning techniques for speech processing. *Inf. Fusion*, 99:101869.

Ji Ming and Danny Crookes (2017). Speech enhancement based on full-sentence correlation and clean speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 25(3):531–543.

Andrew Cameron Morris, Viktoria Maier, and Phil D. Green (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*, pages 2765–2768. ISCA.

Sugan Nagarajan, Satya Sai Srinivas Nettimi, Lakshmi Sutha Kumar, Malaya Kumar Nath, and Aniruddha Kanhe (2020). Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales. *Digit. Signal Process.*, 104:102763.

Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song (2020). A decade survey of transfer learning (2010-2020). *IEEE Trans. Artif. Intell.*, 1(2):151–166.

Jane Oruh and Serestina Viriri (2020). Spectral analysis for automatic speech recognition and enhancement. In *Machine Learning for Networking - Third International Conference, MLN 2020, Paris, France, November 24-26, 2020, Revised Selected Papers*, volume 12629 of *Lecture Notes in Computer Science*, pages 245–254. Springer.

Oyebade K. Oyedotun, Konstantinos Papadopoulos, and Djamila Aouada (2023). A new perspective for understanding generalization gap of deep neural networks trained with large batch sizes. *Appl. Intell.*, 53(12):15621–15637.

Edvin Pakoci, Darko Pekar, Branislav M. Popovic, Milan Secujski, and Vlado Delic (2022). Overcoming data sparsity in automatic transcription of dictated medical findings. In *30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*, pages 454–458. IEEE.

Sinno Jialin Pan and Qiang Yang (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Martha Maria Papadopoulou, Anna Zaretskaya, and Ruslan Mitkov (2021). Benchmarking asr systems based on post-editing effort and error analysis. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 199–207.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Seungcheol Park, Jaehyeon Choi, Sojin Lee, and U Kang (2024). A comprehensive survey of compression algorithms for language models. *arXiv preprint arXiv:2401.15347*.

Theresa Pekarek-Rosin and Stefan Wermter (2023). Replay to remember: Continual layer-specific fine-tuning for german speech recognition. In *Artificial Neural Networks and Machine Learning - ICANN 2023 - 32nd International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, September 26-29, 2023, Proceedings, Part VII*, volume 14260 of *Lecture Notes in Computer Science*, pages 489–500. Springer.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath (2023). Prompting the hidden talent of web-scale speech models for zero-shot task generalization. *CoRR*, abs/2305.11095.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych (2020). Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le (2023). Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658.

Barbara Plank (2011). *Domain adaptation for parsing*. Ph.D. thesis, University of Groningen. Relation: http://www.rug.nl/ Rights: University of Groningen.

Sameer S. Pradhan, Ronald A. Cole, and Wayne H. Ward (2023). My science tutor (myst) - A large corpus of children's conversational speech. *CoRR*, abs/2309.13347.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert (2020). MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2757–2761. ISCA.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black (2021). BABEL: bodies, action and behavior with english labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 722–731. Computer Vision Foundation / IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International*

Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alan Ramponi and Barbara Plank (2020). Neural unsupervised domain adaptation in NLP - A survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6838–6855. International Committee on Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jette (2021). Earnings-21: A practical benchmark for ASR in the wild. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3465–3469. ISCA.

Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Lena Bornholt, Jörn Ostermann, and Ulrike Lüdtke (2022). kidstalc: A corpus of 3- to 11-year-old german children's connected natural speech. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 5160–5164. ISCA.

Martin Russell (2006). The pf-star british english childrens speech corpus. *The Speech Ark Limited*.

Barbara Schuppler, Martine Adda-Decker, and Juan Andres Morales-Cordovilla (2014a). Pronunciation variation in read and conversational austrian german. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 1453–1457. ISCA.

Barbara Schuppler, Martin Hagmueller, Juan Andres Morales-Cordovilla, and Hannes Pessentheiner (2014b). GRASS: the graz corpus of read and spontaneous speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1465–1470. European Language Resources Association (ELRA).

HE Semary, Khamis A Al-Karawi, and Mahmoud M Abdelwahab (2024). Using voice technologies to support disabled people. *Journal of Disability Research*, 3(1).

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah (2020). A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5(1):12.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed (2022). Robust self-supervised audio-visual speech recognition. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2118–2122. ISCA.

Ensheng Shi, Yanlin Wang, Hongyu Zhang, Lun Du, Shi Han, Dongmei Zhang, and Hongbin Sun (2023). Towards efficient fine-tuning of pre-trained code models: An experimental study and beyond. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2023, Seattle, WA, USA, July 17-21, 2023*, pages 39–51. ACM.

Zhongzhi Shi, Bo Zhang, and Fuzhen Zhuang (2012). Improving transfer learning by introspective reasoner. In *Intelligent Information Processing VI - 7th IFIP TC 12 International Conference, IIP 2012, Guilin, China, October 12-15, 2012. Proceedings*, volume 385 of *IFIP Advances in Information and Communication Technology*, pages 28–39. Springer.

Ian Simon, Josh Gardner, Curtis Hawthorne, Ethan Manilow, and Jesse H. Engel (2022). Scaling polyphonic transcription with mixtures of monophonic transcriptions. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, pages 44–51.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha (2023). Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access*, 11:6973–7020.

Matthew E. Taylor and Peter Stone (2007). Cross-domain transfer for reinforcement learning. In *Machine Learning, Proceedings of the Twenty-Fourth International*

Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007, volume 227 of *ACM International Conference Proceeding Series*, pages 879–886. ACM.

Sebastian Thrun and Lorien Pratt (1998). *Learning to Learn: Introduction and Overview*, pages 3–17. Springer US, Boston, MA.

Irfan Ahmed Usmani, Muhammad Tahir Qadri, Razia Zia, Fatma S Alrayes, Oumaima Saidani, and Kia Dashtipour (2023). Interactive effect of learning rate and batch size to implement transfer learning for brain tumor classification. *Electronics*, 12(4):964.

Ayushi Y Vadwala, Krina A Suthar, Yesha A Karmakar, Nirali Pandya, and Bhanubhai Patel (2017). Survey paper on different speech recognition algorithm: challenges and techniques. *Int J Comput Appl*, 175(1):31–36.

Paola Valli (2015). The taus quality dashboard. In *Proceedings of Translating and the Computer 37*.

Thilo von Neumann, Christoph Böddeker, Keisuke Kinoshita, Marc Delcroix, and Reinhold Haeb-Umbach (2022). On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. *CoRR*, abs/2211.16112.

Hsin-Wei Wang, Bi-Cheng Yan, Yi-Cheng Wang, and Berlin Chen (2022). Effective asr error correction leveraging phonetic, semantic information and n-best hypotheses. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 117–122. IEEE.

Olivia Weng (2021). Neural network quantization for efficient inference: A survey. *CoRR*, abs/2112.06126.

Johannes Wirth and René Peinl (2022). Automatic speech recognition in german: A detailed error analysis. In *IEEE International Conference on Omni-layer Intelligent Systems, COINS 2022, Barcelona, Spain, August 1-3, 2022*, pages 1–8. IEEE.

Johannes Wirth and René Peinl (2023). ASR bundestag: A large-scale political debate dataset in german. *CoRR*, abs/2302.06008.

F Wurster, N Cecon-Stabel, T Hansen, J Jaschke, J Köberlein-Neu, MR Okumu, C Rusniok, H Pfaff, and U Karbach (2023). The implementation of an electronic medical record (emr) and its impact on quality of documentation. *European Journal of Public Health*, 33(Supplement_2):ckad160–864.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan (2019). Universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2720–2729. Computer Vision Foundation / IEEE.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu (2019). Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1526–1530. ISCA.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer (2022). Lit: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18102–18112. IEEE.

Haojie Zhang, Ge Li, Jia Li, Zhongjin Zhang, Yuqi Zhu, and Zhi Jin (2022a). Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao (2023). Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shudong Zhang, Haichang Gao, Tianwei Zhang, Yunyi Zhou, and Zihui Wu (2022b). Alleviating robust overfitting of adversarial training with consistency regularization. *CoRR*, abs/2205.11744.

Li Zhou, Suzanne V Blackley, Leigh Kowalski, Raymond Doan, Warren W Acker, Adam B Landman, Evgeni Kontrient, David Mack, Marie Meteer, David W Bates, et al. (2018). Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA network open*, 1(3):e180530–e180530.

Xiaojin Zhu (2008). Semi-supervised learning literature survey. *Comput Sci, University of Wisconsin-Madison*, 2.

Matthias Zuchowski and Aydan Göller (2022). Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. *British Journal of Healthcare Management*, 28(1):30–36.