Improving accented English speech recognition through cross-accent transfer learning

Enes Ihsan Aydogan

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2025

Abstract

This study explores ways to improve the recognition of accented English in automatic speech recognition (ASR) systems. Despite significant progress in ASR technology, accented speech recognition remains challenging due to limited targeted research, a lack of diverse datasets, and variations in pronunciation and speaking styles of English speakers. Our research investigates the trade-offs between using mixed accented training datasets versus focused, single-accent datasets for fine-tuning. We also discuss relevant topics in accented speech recognition, such as categorising speech samples based on their features, comparing the effects of language models and fine-tuning on performance, and analysing the outcomes of fine-tuning in different contexts. We use Wav2Vec2 2.0 model, with VoxPopuli and EdAcc datasets to examine these questions. Our findings demonstrate that fine-tuning with limited data can lead to considerable improvements in recognising in-domain speech. However, we also highlight the limitations and challenges associated with transferring these learnings across different domains. These methods have the potential to create more robust and effective ASR systems, benefiting English speakers with diverse accents and promoting inclusivity in ASR technology.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Enes Ihsan Aydogan*)

Acknowledgements

Mum, Dad, and Nunu, for their support, encouragement, and patience.

Roksana, my dear sister Sevde, Ismail V, Ma'arij and Wisham, for their friendship and the memories they blessed with their presence.

James Garforth, for his continuous guidance and valuable insights throughout.

Edinburgh, Glasgow, London. My homes.

Lastly the main library for the countless days I have spent in there this year. Also the same as last year pretty much because I am in a rush.

Table of Contents

Chapter 1

Introduction

Automatic speech recognition (ASR) is the process of converting human speech into machine- or human-readable formats, such as embeddings or text, to enable human-human or human-computer communication [Juang and Rabiner, 2005]. Despite significant developments, there is still a performance gap between marginalised demographics and accents persist [Markl and McNulty, 2022, Jahan et al., 2025].

1.1 Motivations

This study aims to investigate ways of bridging the performance gap in accented English speech recognition on a range of L2 English accents. Much of the work done on the matter has focused on regional native accents. We aim to include a variety of European non-native accents, explore how much improvement can be achieved, whether it is possible to leverage more high-resource accents to improve other accents with transfer learning, feasibility of clustering speech based on features rather than pre-determined labels, comparisons of accuracy gains with potential improvements from language model additions to models, investigating the cross-domain generalisation of the fine tuning across accents.

1.2 Overview

In this dissertation, we investigate ways of improving the performance of ASR systems in foreign-accented English speech (English spoken by non-native speakers). I conduct experiments to examine the current performance of some contemporary models. These experiments explore how accented English speech recognition can be improved using little training data, the relationship and potential transferability of learning between phonetically or linguistically similar accents, the impact of language models on ASR systems, and the cross-domain applicability of learning. To our knowledge, there are no studies that look into mixed-accent training, accent-specific training, and cross-accent improvements between accents simultaneously.

We explore specific research questions about accented English speech recognition, with

each experiment addressing one or multiple questions.

1.2.1 Experiments

The first two experiments are on fine-tuning and testing a baseline model using various splits with an accented dataset. In the first experiment, a combination of accented samples are selected into a single training set, where each accent gets a similar number of samples. Thus, no one accent is overrepresented in the combined set. The fine-tuned model is then evaluated using the test split, which contains the same variety of accents. This experiment explores how vital variety is in the training data for accent improvements to generalise and cover a wide range of accents.

The second set of experiments focuses on different train-test splits, training on a single accent while evaluating across all accents. This aims to identify cross-accent improvements gained through fine-tuning. It complements the first experiment by assessing whether to focus on specific accents for better performance in accented ASR with limited data. If combined training struggles with diverse accents, concentrating on single accents may be a more effective strategy for resource gathering and training.

The third set of experiments builds on the previous ones but addresses the limitations of the native-language heuristic for categorising accents. Instead of predefined categories, we grouped speech samples based on feature similarity, creating unnamed clusters. This approach aims to improve test data by clustering accents and fine-tuning models on subsets of these clusters, allowing us to train and evaluate using data most similar to the model's training samples.

The fourth setup explores how a language model added to an ASR system may impact the performance in accented speech recognition and whether it can match the improvements gained by fine-tuning. As language models are a common component in ASR systems, it raises the question of whether the improvements we achieve from the first three experiments focusing on improving the acoustic model performance can be achieved with a language model head instead.

Lastly, our final experiment explores the cross-domain transferability of the training. We use the best-performing model from the first two experiments. The fine-tuned model is then evaluated by its performance on the test set of the original dataset along with another dataset containing speech samples with a different context, register, and setting. The purpose of this experiment is to examine the performance of the fine-tuning on data from different and unseen domains.

1.2.2 Research Questions

These experiments aim to provide a clear and comprehensive understanding of how contemporary automatic speech recognition systems can be improved for accented English speech. The list of topics and research questions I intend to address are listed below:

1. How much improvement to accented speech recognition can be realised with a dataset limited in size but containing a wide range of accents?

2. If we focus on a single accent in a targeted manner in the training dataset instead of a wide range of accents, how does this approach improve the model's performance on accents similar to the training accent? Is there a potential for transfer learning between accents?
3. As accent labels based on native language may not necessarily be representative in all cases, is it possible to create categories for speech samples by looking at the speech and its features only? creating abstract categories instead of labels to collate samples by feature similarity?
4. Does a language model on an ASR model improve the accuracy as much as our fine-tunings from the previous experiments?
5. Do the improvements from fine-tuning on one dataset impact the performance on another accented dataset containing speech in a different context?

These serve as an overall guide to the structure of the rest of the report, including background reading, explanations of methodology, and analysis of the results.

1.2.3 Contributions

Our findings demonstrate a strong potential for improving accented English ASR, even when using limited data with a high variety in accents. With fine-tuning on small, accent-mixed training sets, we observe improvements—often more than 25%— in word error rates. When adopting a more targeted approach in fine-tuning, supplying the same amount of data but focused on a single accent, we achieve even improvements between 23% and 47% across all accents, with most improving by more than 34%.

Accent-agnostic clustering and attempts at abstracting away the accent labels yielded inconclusive results. MFCC-based clustering seems to include too random of a mix of accents, but Wav2Vec2-based clusters separated worse-performing summary

Our comparison of improvements from language model addition and fine-tuned model results clearly shows that fine-tuning improves accuracy far more than language model addition. Lastly, our results indicate that domain mismatch is a persistent issue across fine-tuned models, with potentially fine-tuning in a niche domain causing degraded performance from the baseline model for some accents.

The main contributions of this work are: (1) the scale and breadth of the accents explored in the fine-tuning and evaluation; (2) a detailed comparison between mixed-accent and accent-targeted fine-tuning performance; and (3) an investigation into the interaction between cross-accent and cross-domain interactions, revealing patterns, symmetries the in lack of transfer between domains, and possible degradation to baseline performance for niche domain-adaptations.

1.3 Structure of the Report

The report is split into several chapters.

Chapter 2 - Related Work provides an overview of the field of automatic speech recognition, accented speech recognition and why is it a problem, and previous work done on some of the relevant topics.

Chapter 3 - Methodology explains the data processing, experiment setup steps, and evaluation methods in depth.

Chapter 4 - Results details extensive investigation and interpretation of the results from the experiments. Overview of what patterns are observed, and for the experiments with inconclusive results, potential causes for the inconclusive results.

Chapter 5 - Discussions provides a comprehensive overview of the study's limitations, further work and expansion possibilities, and final remarks on the contributions of this project.

1.4 Following on From Last Year

My UG4 project was on a different field and question. Last year, my dissertation was titled "*Detecting and combatting misinformation online during crises: a case study of COVID-19*." It was a natural language processing project in which I created a pipeline to identify and provide fact-checked sources for selected claims about the COVID-19 virus and the pandemic. The project aimed to combat the prevalent problem of misinformation online by suggesting a potential tool to allow users to quickly identify and get more information about misinformative claims.

Around the same time I was doing my write-up, I was also working on another significant coursework for the Machine Learning Practical course. The final project for the course was a broad quest to take some problem and come up with a machine learning focused solution to it. After some brainstorming, we came up with the idea of looking into accented English speech recognition. In the MLP project, we broadly focused on the transferability of learning between accents only, which is described in this paper as the second experiment in the section 1.2.1, corresponding to research question three. I greatly enjoyed working on the project, and by the time I finished, I had more ideas on how it could be improved and what kinds of additions could be made. As there were more tangents to explore, I decided to turn the MLP project into my dissertation topic this year.

Due to teammates' personal constraints and special circumstances, I did all the practical work on setting up the experiments, working with HuggingFace API, collating the data, and the experiment results. The teammates helped in the final write-up, and the presentation of the data in figures and some preliminary analysis of another dataset (Speech Accented Archive) that we did not use in the final paper as it diverged from the story we wanted to tell in the report.

The experiment setup I created last year had certain problems that we only realised the severity of later. The most prominent issue was the potential risk of selecting an inconsistent amount of data between experiments, which could significantly impact the performance after fine-tuning. This paper thus aims to correct this issue and explore the other research questions I did not have the time for last year.

Last year's work has been beneficial in assisting my understanding and competency with the tools I have used for this project, such as the HuggingFace Trainer API and dealing with speech data. But a final important note is that all the work presented in this paper was done by me independently. None of the code, implementation, or work from last year was used in any capacity. Throughout this report, I use "we" and "our" to denote my personal work, as this is the format I am more comfortable writing in. This is merely a stylistic choice I took to express my work more comfortably throughout the report.

Chapter 2

Related Work

2.1 Automatic Speech Recognition

Automatic speech recognition (ASR) has developed significantly from early methods using rules-based and statistical modelling approaches to contemporary state-of-the-art deep learning based approaches [Yu and Deng, 2016]. Over time, the ability of these models to deal with a variety of inputs in different domains of speech with varying levels of quality has significantly increased. There is substantial ongoing research and interest in improving the quality of these models for recognising different languages with higher accuracy, transcribing them into text or converting them to machine-readable formats such as vector embeddings for further processing, categorisation, or transcription [Yu and Deng, 2016].

Early ASR systems mainly relied on statistical models such as Hidden Markov Models with Gaussian Mixture Models and rules-based approaches from the phonetics and morphology of the target languages. Although these methods can be effective with limited vocabulary and speaker-dependent accuracy, they struggle with scalability and successfully dealing with a wide variety of speech, such as background noises and accents [Juang and Rabiner, 2005]. The rising prominence of machine learning, particularly sophisticated deep learning methods using neural networks and transformer models, has impacted the performance, robustness, and accuracy of automatic speech recognition [Ngueajio and Washington, 2022]. State-of-the-art models based on deep learning and end-to-end transformers, such as Wav2Vec2, Whisper, and Conformer architectures, leverage self-supervision with large amounts of data to improve performance across diverse languages, speakers, domains, and accents [Fuckner et al., 2023].

Modern ASR systems in use contain several key components in their pipeline: feature extraction, acoustic models, language models, and decoding. Feature extraction converts raw speech features into more informative representations, such as spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs). Acoustic models map these features to phonetic units. Language models leverage linguistic context and information to improve accuracy, and decoding combines the acoustic and language models into a final transcription [Malik et al., 2021].

End-to-end ASR models, such as those based on Connectionist Temporal Classification (CTC) and sequence-to-sequence approaches with attention mechanisms, have further streamlined ASR pipelines by directly mapping speech input to text output. These streamlined approaches have shown significant improvements in performance and robustness compared to more traditional methods and are widely adopted in ASR research and development [Das et al., 2018].

Despite its strengths, some limitations persist in making these ASR systems robust, adaptable and inclusive. Currently, data for training models are available only for a handful of languages out of approximately 6500 world languages [Malik et al., 2021]. Automatic speech recognition is also, much like other machine learning tasks, highly data-dependent, and the availability of large-scale datasets plays a crucial role in the developments. Current systems also suffer from various biases on gender, age, and accent [Fuckner et al., 2023, Jahan et al., 2025].

2.2 Accented Speech Recognition

Contemporary state-of-the-art ASR systems perform well in native English speech, particularly with high-resource native accents, due to the biases in the pre-training data. As a result, these systems perform optimally when processing clear speech spoken in standardised accents, and their accuracy significantly degrades when presented with non-standard accents, including regional native English dialects that deviate from the dominant accents or non-native English spoken by L2 speakers with foreign accents [Malik et al., 2021]. This performance gap is a key limitation in the accessibility of contemporary ASR systems to non-dominant accented speakers [Markl and McNulty, 2022]. This is especially a prominent problem for English, as most English speakers are not native speakers [Crystal, 2003].

Everyone has an accent, even speakers of dominant accents of a language who are often thought to have "no accent" [Markl and Lai, 2023]. For the remainder of this paper, the term "accented speakers" denotes people who speak a non-standard accent, either a native but regional dialect that diverges from standard local variations or L2 English speakers with foreign accents.

2.2.1 Negative consequences of suboptimal accented ASR

This performance gap can have real-life consequences for many people with non-standard accents, a central pain point in human-computer interaction for those who do not or cannot imitate the standard accents to be understood by ASR systems. For people with marginalised accents, this can be significantly detrimental to their experience with ASR systems. This performance gap's consequences can be far reaching, preventing accented speakers from interacting with commercial products with ease, impacting their ability to conduct simple tasks with their devices that may have ASR systems built-in, such as cars or electronic devices, and in severe cases even negatively impact the quality of service they receive in healthcare and damage their employment opportunities [Markl and McNulty, 2022]. These negative consequences often worsen the status of these marginalised communities, who already face negative biases in various forms in their

day-to-day life, and the lack of consideration about this matter only propagates the problem [Markl and McNulty, 2022].

Accented speakers may struggle interacting with commercial products and face difficulties performing tasks on devices with built-in ASR systems (such as cars or electronic devices). The consequences of relatively worse performance for commercial products may not be critical in the users' lives, but they can cause a sense of exclusion and frustration. For example, persons with accents may find it harder to interact with voice-activated assistants such as Siri or Alexa, and their requests may not be accurately parsed by these systems [Ngueajio and Washington, 2022].

ASR systems are also used in healthcare settings for tasks such as transcribing clinical records, controlling medical equipment, and automatic translation systems [Chiu et al., 2017, Blackley et al., 2019]. These systems' failure to recognise accented speech accurately can lead to inaccurate medical records and other harmful errors in treatment, which can be very detrimental to the health of the patients [Olatunji et al., 2023].

Another critical utilisation of ASR systems is in human resources and employment. Many companies use automated tools in their hiring processes, online interviews and recorded answers to these questions [Hickman et al., 2024]. Ensuring consistent and high accuracy for the ASR systems used in this context is crucial for the fairness of the hiring process and preventing any unintended biases in rejecting candidates due to errors or shortcomings of the ASR systems. Bias against accented speakers here can have a real economic impact for those who do not conform to the norms expected by ASR systems for optimal accuracy [Markl and McNulty, 2022].

In education, ASR systems are also used in various settings, such as transcribing lecture materials for better accessibility and assisting in language learning [Van Doremalen et al., 2016]. Failing to adequately transcribe accented speech from lectures or other materials can impact the student experience.

Lastly and most crucially, the impacts of these disparities can be long-lasting. If unaddressed, these biases and the lack of attention given to them can perpetuate inequalities, reinforcing existing biases against marginalised communities and contributing to their exclusion from opportunities in the social and economic systems through unfair hiring process setups or problematic medical record-keeping. Thus, these accuracy gaps exacerbate the marginalisation of these communities now and potentially in the future by setting up a negative precedence if not addressed appropriately.

2.3 Models & Datasets

2.3.1 Datasets

Numerous large datasets focus on native speech in different languages. Some early datasets, like the Wall Street Journal [Paul and Baker, 1992] dataset containing over 400 hours of speech, have enabled ASR research for decades. Currently, many comprehensive datasets for English speech are being used in ASR research. LibriVox [Kearns, 2014] contains over 6,000 full-text audiobooks in 48 different languages, and

LibriSpeech [Panayotov et al., 2015], as part of this incentive, contains 1000 hours of recordings for audiobooks in English. Gigaspeech [Chen et al., 2021] contains 10k hours of high-quality speech with labels, along with 40k hours of speech without labels. The CommonVoice [Ardila et al., 2019] dataset contains over 2.5k hours of crowdsourced audio in 38 different languages. TED-LIUM [Hernandez et al., 2018] is another dataset of cleaned and aligned TED talks containing 452 hours of speech. The VoxLingua107 [Valk and Alumäe, 2021] contains over 6.6k speech samples in 107 different languages, accompanied by 1609 verified utterances. CMU-Arctic [Kominek and Black, 2004] includes carefully curated 1200 phonetically balanced English utterances mainly designed for speech synthesis research. These datasets provide a solid foundation for developing ASR systems that can accurately recognise standard English speech.

Numerous datasets have been introduced to classify or improve accented English speech. CommonAccent [Zuluaga-Gomez et al., 2023] is a subset of the CommonVoice dataset, containing mainly native accents. AccentDB [Ahamad et al., 2020] contains samples of four Indian-English accents and a compilation of four native-English and metropolitan Indian-English accents. Speech Accent Archive (SAA) [Weinberger and Kunath, 2011] dataset features parallel speech samples by people from 177 countries. Speakers read the same paragraph, which makes it ideal for studying accent variations. Similarly to SAA, the IDEA (International Dialects of English Archive) [Persley, 2013] contains roughly 1700 samples by people from 135 countries and territories reciting the same text. CSLU Foreign Accented English [Lander] includes speech by native speakers of 22 different native languages. The Interspeech 2020 Accented English Speech Recognition Challenge (AESRC) [Shi et al., 2021] dataset provides eight sets of accented data from different countries, aiming to promote research in accent speech recognition. The L2Arctic [Zhao et al., 2018] dataset is intended for research in voice conversion, accent conversion, and mispronunciation detection, featuring 10 non-native accents. VoxPopuli [Wang et al., 2021] contains a substantial amount of native, as well as 29 hours of accented speech by native speakers of 16 European languages. EdAcc [Sanabria et al., 2023] contains almost 40 hours of dyadic call conversations between friends in a range of native and non-native accents. The GLOBE [Wang et al., 2024] dataset contains 535 hours of speech, mainly aimed at text-to-speech research. Lastly, the VCTK dataset mentioned in [Inoue et al., 2025] includes speech from 109 native speakers from the US and the UK.

Many of these datasets either focus on native-accented English speech (CommonVoice, VCTK), contain little variation in their speech samples (SAA, IDEA) or contain few non-native accents (AccentDB). CSLU, EdAcc, VoxPopuli and GLOBE provide wide-ranging accents with a considerable sample size.

2.3.2 Pre-trained Models

Along with the various large datasets containing native, regional, and non-native English speech, some established large ASR models are available and widely used in research. Many contemporary state-of-the-art models use a combination of convolutional neural networks with transformer architecture [Vaswani et al., 2017] to achieve high accuracy

and robust speech recognition capabilities. This combination is a powerful tool in enabling ASR models to perform well.

Wav2Vec2 by Baevski et al. [2020] is a self-supervised learning framework for speech recognition by Facebook AI. Based on a combination of multi-layer convolutional neural network and transformer architecture, it leverages large amounts of unlabelled data to learn meaningful representations during pre-training and some labelled data with a Connectionist Temporal Classification loss [Graves et al., 2006] to be used for downstream tasks such as classification or speech recognition. Whisper [Radford et al., 2023] is a model developed by OpenAI. It is trained on a large and diverse dataset of speech data in multiple languages. It uses a transformer-based architecture and can perform speech recognition and translation tasks. The Conformer [Gulati et al., 2020] model by Google introduces a convolution-augmented transformer model for speech recognition. Some other models include QuartzNet [Kriman et al., 2020] by NVidia (based on their previous model Jasper [Li et al., 2019]), Deep Speech 2 [Amodei et al., 2016] by Baidu Research. Since these models are often introduced by corporations, their pre-training data is often not publicly available.

Many of these models perform well with native speakers with dominant accents, but there is a substantial increase in the word error rate if they are presented with heavily accented speech, native or non-native [Sanabria et al., 2023, Fuckner et al., 2023]. This is due to the training models

2.4 Speech Feature Representations

This part is motivated by the fact that labels on speech can lack nuance. Thus, we aim to look into ways of representing speech by its features. There are various methods for extracting features of speech. Mel-Frequency Cepstral Coefficients (MFCCs) are the most commonly used. They filter frequencies to represent what humans hear. Another feature is spectrograms, which are used to understand the quality of the sound. Chroma features are used for analysing and processing musical data. The Spectral centroid calculates the weighted mean of the amplitude of frequency, and spectral roll-off calculates the frequency below a certain percentage of the total frequency of the spectrum [Singh et al., 2020].

Self-supervised speech representations have also been highly studied recently. Speech2Vec [Chung and Glass, 2018] proposes a novel deep neural network architecture for learning fixed-length vector representations, similar to Word2Vec in NLP [Church, 2017]. Wav2Vec2 embeddings, the output of the Wav2Vec2 model at the last layer, can also be used as vector representations of speech samples. Pepino et al. [2021] uses these embeddings to create a system to recognise emotions with simple neural networks.

These features can be insightful into the speech samples, including about the phonetic and accent data of the speech.

2.5 Transferability of Learning

Transfer learning is a critical area of research in ASR. Broadly speaking, it can be described as leveraging knowledge from a pre-trained model on a large dataset to improve performance on a smaller, either low-resource or domain-specific dataset. This approach has various use cases in accented speech recognition, where models pre-trained in standard English are fine-tuned on accented speech data. This method helps adapt the model to the nuances of the accent, thus improving accuracy [Sullivan et al., 2022]. Conceptually, this is similar to cross-domain adaptation in section 2.7 and generalisation. Many papers approach this problem as such, but in the context of accented ASR, we use it to transfer learning between accents only in this paper.

In our specific case in this work, we also look into the transfer learning of the fine-tuned model into different accents than the fine-tuning accent. Native speakers of certain languages—such as Czech and Polish—have similar accents in English due to the close connection between their native languages linguistically and phonetically [Slámová, 2018]. This presents a valuable opportunity in leveraging not only the pre-training data but also the fine-tuning data between accents.

Previous work done on this matter mostly focuses on improving accents individually [Hinsvark et al., 2021]. There is some work into investigating the performance changes between accents for fine-tuned models, but they often focus on native accents or lack diversity in their accent variation [Ahamad et al., 2020, Sullivan et al., 2022].

Some researchers introduce novel and ingenious techniques to improve accented ASR accuracy. Gu et al. [2024] proposes a framework to capture similarities between source and target accents to improve cross-accent speech recognition. Another interesting study of leveraging the native-language background of English speakers was done by Kumar et al. [2023]. The study introduces unlabelled native language data to the model during the pre-training phase to learn self-supervised representations. The pre-trained model is then fine-tuned using limited labelled English data in the accent. Applicable to transfer learning and domain generalisation, Finn et al. [2017] introduces meta-learning, a potent transfer-learning technique particularly effective in dealing with low-resource-related challenges. In meta-learning, a model is trained across multiple tasks, allowing the model to acquire 'meta' parameters, which can be adaptable to various other tasks.

2.6 Language Models in ASR

Automatic speech recognition used to have two different fundamental fields, isolated word recognition and continuous speech recognition [Scagliola, 1985]. The inclusion of language models (LMs) in ASR and the incorporation of context into the task of transcribing audio input made continuous speech recognition viable [Liu et al., 2024a]. LMs played an essential role in ASR systems by providing contextual information to improve accuracy, utilising statistical data from vast text corpora to predict the likelihood of word sequences, and integrating with the acoustic models to leverage our understanding of text, grammar, and context. They also assisted ASR systems in enabling disambiguating similar-sounding words and phrases. They reduce the word

error rate by replacing unlikely words transcribed by the acoustic model with more likely words in a given context. However, with the rising prominence of neural end-to-end models, LMs have fallen out of use Liu et al. [2024a].

There are different language models: n-grams, neural network-based ones, and transformer models. N-gram models predict the probability of a sequence of length n-1 based on some training corpora. They are simple and efficient but suffer from data sparsity and lack context beyond the previous n-1 tokens [Jelinek, 1998]. One common n-gram model used in ASR systems is KenLM [Heafield, 2011], due to its efficient and fast search algorithm. Neural language models are more sophisticated LMs that can capture long-range dependencies. They outperform n-gram models by learning complex patterns [Mikolov et al., 2010, Bahdanau et al., 2014]. Transformer-based models [Vaswani et al., 2017] such as BERT and its variants have revolutionised the natural language processing field by introducing self-attention mechanisms to effectively capture contextual information.

Language models can be integrated into ASR systems in different ways. Shallow fusion models combine acoustic models and language models during the decoding process, where the most likely sequence predictions from the acoustic model are selected by the language model based on linguistic context. It is an intuitively straightforward approach widely used in speech recognition. Deep fusion models involve training acoustic and language models jointly to optimise their combined performance [Gülçehre et al., 2015]. Cold fusion models use a language model to guide the training of the acoustic model, allowing it to benefit from the linguistic context during the training process [Sriram et al., 2017].

The language model addition to ASR systems increases accuracy most of the time [Toshniwal et al., 2018]. Some previous studies have used n-gram models, [Hirsimaki et al., 2009, Pohl and Ziółko, 2013, Habeeb et al., 2021, Kumar and Niranjan, 2024], and optimised search algorithms for more efficient n-grams [Heafield, 2011], uni- and bi-directional RNNs and LSTMs [Arisoy et al., 2015, Lam et al., 2019, Irie et al., 2016], and lastly transformer-based LLMs [Min and Wang, 2023, Chiu and Chen, 2021, Futami et al., 2020] with varying degrees of success. For early LLMs such as BERT, some papers report increased accuracy [Chiu and Chen, 2021, Futami et al., 2020], but there are also reports of decreased accuracy [Min and Wang, 2023].

With transformer-based models, Min and Wang [2023] reports the inclusion of LLMs increases the word error rate, while [Futami et al., 2020] and Chiu and Chen [2021] report increased accuracy.

2.7 Domain Generalisation and Adaptation

Machine learning models' performance can often fail to handle significant changes between training and test data. Shifts in the data domain for ASR systems, such as between technical monologues and casual conversations or even accent changes in the samples, can pose a problem to the adaptability of the models [Sun et al., 2016]. This can be seen as a domain generalisation problem, where the target data may come from a different domain [Blanchard et al., 2011, Muandet et al., 2013].

This issue of domain changes can be tackled in various ways, such as using meta-learning methods [Li et al., 2018, Finn et al., 2017], introducing domain-shift during training, augmenting data with domain synthesis [Zhou et al., 2022], or fine-tuning pre-trained models with domain-specific data [Luo et al., 2021]. Such methods are used in a variety of fields within machine learning research, such as computer vision, natural language processing, medical imaging, and speech processing [Zhou et al., 2022].

Paraskevopoulos et al. [2023] and Zhou et al. [2023], along with other cited papers in this section, suggest that pre-trained models often fail to perform as accurately in out-of-domain speech samples and cause higher WERs.

Chapter 3

Methodology

Overall, we ran five sets of experiments using the Wav2Vec2 model [Baevski et al., 2020]. Each experiment addresses one research question mentioned in section 1.2.2. We used the accented subset of the *VoxPopuli* [Wang et al., 2021] dataset in most experiments. This dataset consists of L2-English speech samples by native speakers of 16 different European languages, obtained from European Parliament speeches. It provides a wide range of non-native accents, with some from the same language families (Polish, Czech, Slovak, etc.) to examine how feasible transferability of learning between phonetically and linguistically similar languages is. For the final experiments where the cross-domain transferability is explored, we utilised a selected subset from the *EdAcc* [Sanabria et al., 2023] dataset. Lastly, HuggingFace Trainer API was used in data processing, training, and evaluating the models [Jain, 2022].

3.1 Dataset Preparation

The VoxPopuli dataset is present in the HuggingFace platform and usable with their Datasets package in Python. We performed some exploratory analysis of the data, such as counting the number of speech samples, words, mean, median, standard deviation between samples, lengths of longest and shortest samples, roughly how much time in minutes do they correspond to at a rate of 160 words per minute [Tauroza and Allison, 1990]. Additionally, 500 samples from the native English set of VoxPopuli were also included in the dataset to compare model performance in native and non-native English performance.

We have performed some pre-processing in the selection process for the training datasets to ensure that each accent gets an equal amount of representation with samples of average length relative to the accent subset. This aims to avoid very short sentences in the training. The data pre-processing steps are explained in further detail in each subsection below.

The EdAcc dataset was also used in the last experiment investigating the transfer of learning across different domains and registers. EdAcc contained many short samples, significantly increasing the perceived word error rate. Some filtering was done to

Accent	Count	Length	Mean	Median	S.D.	Max.	Mins
Slovenian	66	1825	28	24	16	80	11
Croatian	127	3391	27	23	23	238	21
Lithuanian	178	4620	26	23	22	258	28
Estonian	289	8364	29	25	20	172	52
Italian	360	9434	26	23	16	100	58
Slovak	409	10720	26	22	16	125	67
Spanish	419	11434	27	23	16	98	71
Romanian	485	14335	30	23	22	167	89
Finnish	590	15317	26	23	17	193	95
Hungarian	676	19495	29	25	17	111	121
French	737	19701	27	24	19	339	123
Polish	852	23874	28	25	16	130	149
Czech	974	26922	28	25	15	98	168
German	1074	29045	27	23	19	238	181
Dutch	1151	30039	26	23	17	196	187
English	500	11732	23	21	14	80	73

Table 3.1: Overview of the VoxPopuli accented dataset

include only meaningful samples in our training and evaluation.

The overview of datasets VoxPopuli and EdAcc are shown in the tables 3.1 and 3.2, respectively. Due to the way samples are spread in the dataset, with some containing as little as one word, we took some steps to ensure a meaningful spread of samples for training. In essence, this ensures that training samples are of similar lengths and do not include very short or one-word samples. More detailed steps for each set of experiments are detailed in the subsections.

3.2 Model

The Wav2Vec2 base model by Facebook, which was pre-trained on 960 hours of Librispeech data, was used for all experiments [Baevski et al., 2020]. Although some other models perform better in general, such as Whisper by OpenAI, the main issue with using this model is the lack of transparency in their training data, which may include the datasets used in this setup [Radford et al., 2023]. To avoid this, we have decided to use a model with pre-training data that is publicly available and focus on evaluating the improvements from the baselines that can be achieved and the relationship between improvements of different accents.

3.3 Hyperparameters

First, we run some hyperparameter setup experiments on the combined train-test. The decision to use a combined train-test split (explained in 3.4.1) for the hyperparameters

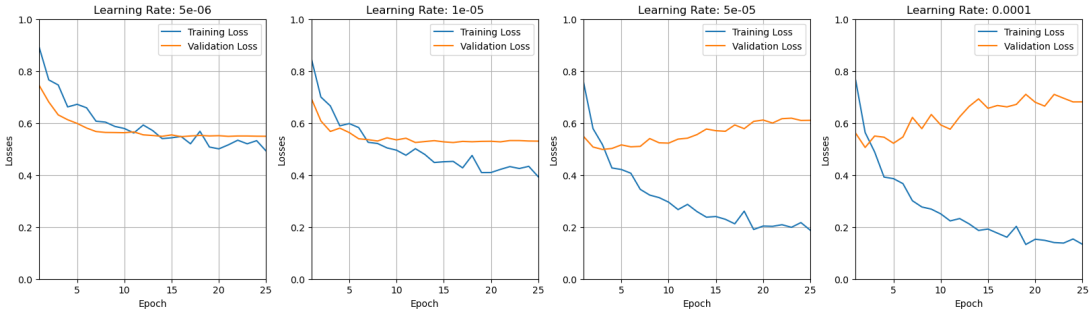

Figure 3.1: Training and validation loss for hyperparameter selection experiments.

selection setup was made to avoid favouring any one accent over others. The number of training epochs and the learning rate were chosen with these experiments.

We run the trainer on the combined train and test split four times with the learning rate values of 1e-4, 5e-5, 1e-5, and 5e-6 for 25 epochs each to determine where the validation error stops changing significantly and which learning rate yields the lowest validation loss overall. Results from these experiments are shown in figure 3.1. Accordingly, we use 1e-5 learning rate, and 12 epochs of training for subsequent experiments to keep a consistent setup across all experiments. This ensures that the only changes we make to the experiments are the provided data and other constants are static across all runs.

A weight decay of 0.005 is also added to prevent overfitting and improve model generalisation. All experiments are run on Google Colab using a T4 GPU with a batch size of 4 (decided due to computational constraints, as higher batch sizes caused the runs to crash due to memory errors).

Below are the training arguments used for the experiments.

```
training_args = TrainingArguments(
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    learning_rate=0.00001,
    weight_decay=0.005,
    num_train_epochs=12,
    gradient_checkpointing=True,
```

3.4 Experiments

3.4.1 Combined Experiments

The main purpose of these experiments is to examine the improvements that can be gained by fine-tuning the model with a wide range of accents. Each accent contains a similar number of speech samples and words in the training set, so no one accent dominates the training. This will give us an idea of whether we can realise significant improvements with a small amount of training data from numerous different accents.

The first step is to separate the data into train and test samples. For this set of ex-

periments, there are two core qualities the data must conform to to make the results meaningful and useful for the aims of this experiment and comparison with future experiments. These are:

1. Each accent must have equal or similar representation. This ensures that no one accent overpowers other accents when evaluating the performance of the fine-tuned model across a wide range of samples from various accents.
2. Enough data is needed for the model to learn through training.

In total, 8000 words are allocated to training datasets. This number is selected as it allows us to have as many accent-specific experiments as possible for the second set of experiments (see section 3.4.2) without severely restricting the amount of data (see table 3.1). This amount is also set uniformly across experiments to ensure consistency across experiments for exploring the research questions, focusing only on the variety in the accents and training data instead of any possible discrepancies in the total amount of training data. For the combined experiments here, each accent was allocated an equal portion of the 8000-word limit, so about 533 words maximum per accent.

To avoid very short samples containing single words or phrases, we sort the samples by accent and how much each sample's length differs from the average length for its accent. This prioritises the samples whose length is closer to the average word count of the accent in the training set. Once the per-accent 533-limit is reached, the rest of the data is allocated to the test set, and we move on to other accents.

After this step, three experiments are run:

1. No validation split: This feeds all 8000 words into the training with no validation split.
2. Holdout - 75/25 Train/Val split: 6000 words used for training and 2000 words for validation.
3. KFolds: 75/25 split run four times with four possible train/validation rotations. This ensures that we account for possible different initialisations and train/validation splits, as these types of training can be very data-dependent.

The results from these runs can also provide valuable insights when compared with the experiments in Section 3.4.2. This comparison helps explore the trade-offs between concentrating on single accents during data collection and focusing on gathering more diverse data. The goal is to ensure that the trained models are robust and capable of handling a variety of accents effectively.

3.4.2 Accent-Specific Experiments

Similarly to combined experiments from section 3.4.1, for the accent-specific experiments and exploring the transferability of learning between accents, we prepare the data into the correct format for training and evaluating. The VoxPopuli dataset is split into a dataset dictionary by available accents. Each accent is further processed to extract 8000 words into the accent's training set and the rest into the test set. As described in the section above, the data is also sorted to prioritise the samples containing the closest

number of words to the average no-of-words. This ensures that the training data does not contain too many very short samples containing only a few words or too long that they take up space and reduce sample diversity in the dataset.

After creating train-test splits for each accent, we fine-tune the baseline Wav2Vec2 model, which has been pre-trained on 960 hours of Librispeech data, using the training set specific to each accent. We then evaluate the fine-tuned model across all accents. For the selected training accent, we use only the test set, while for the other accents, we utilise all available data. This approach ensures that the evaluation data has never been seen by the model before. Since the training on each accent relies solely on that accent's training set, we can use the train and test sets of other accents without issues, as these sets are not presented to the model during fine-tuning.

We hypothesise that the most pronounced improvements are expected on the training accent's test data. We further hypothesise that speech from speakers with similar native languages, such as Czech-Polish-Slovak, would likely show more improvement in each other's speech than they would with other accents due to the phonetic and linguistic similarity of these languages and the English accents of their speakers [Slámová, 2018].

The obtained results should preferably be valuable in determining the feasibility of cross-accent learning of ASR models with fine-tuning on some accented speech. This can be useful in numerous ways, including but not limited to guiding data practitioners in collecting and gathering data, leveraging high-resource accents for improving low-resource accented speech recognition, and potentially further investigating similarities between accents and speech from people with different accents.

One concern with this approach is that the model adapts to the domain of speech instead of to the accents, in this case, monologues in the European Parliament. We address this concern by fine-tuning the model on native English speech, using a similar processing and train-test split as other experiments. We hypothesise that if the fine-tuning causes the model to adapt to the speech domain instead of the accent, the model fine-tuned with native English speech will improve the performance across all accents, similar to other experiments with different accented training data. If the model learns the accent distinctions instead, native English fine-tuning will not cause significant improvements.

3.4.3 Accent-Agnostic Experiments

The above two experiments and the performance improvements we obtain rely on an assumption: the native language label allocated to each sample correctly represents the accent of the speaker in that sample. This, however, is not always true. Only taking the native language of a speaker into account when determining their accent ignores many other factors, such as education level and personal background of the person [Sanabria et al., 2023]. The experiments in this section aim to address this issue.

To eliminate label bias, we aim to look at the features of the audio samples alone and categorise them accordingly. This approach tries to abstract the accent classes and cluster samples according to their similarity rather than relying on labels like the speakers' native languages, which lack granularity in representing the speakers' access. In essence, this is similar to accent recognition, but instead of evaluating the success

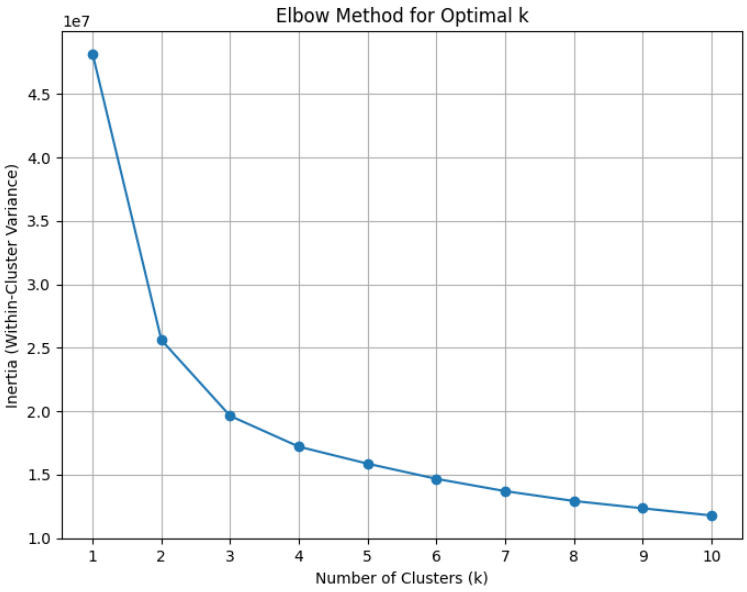

Figure 3.2: Elbow plot for selecting the number of clusters for K-Means.

of the clustering by looking at whether it correctly classifies samples, we measure it through overall improvements to word error rates.

For this, we look into two ways of extracting features or identifying computational representations of the audio samples: Mel-Frequency Cepstrum Coefficients (MFCC), and Wav2Vec2 audio embeddings. Singh et al. [2020] conducts an experiment in order to find the best performing feature extraction method among the ones mentioned in 2.4, and MFCCs yield the best performing model trained. Thus we choose MFCCs to cluster samples into unnamed groups. Wav2Vec2 embeddings can also be used for various tasks [Pepino et al., 2021].

Principal component analysis (PCA) is a method to reduce dimensionality of a vector while retaining as much information as possible. It achieves this by transforming original variables into new variables called principal components [Wold et al., 1987]. As the data we obtain from MFCC extraction and Wav2Vec2 embeddings are highly dimensional, we perform PCA on the results to collate the results into unnamed groups. As there were some outlier samples in the dataset that had their own tiny clusters at the far peripheries of the components, some filtering on the data was done and all samples containing less than 5 words were removed.

We then reiterate the process of gathering train and test splits in these clusters. We hypothesise that relatively similar accents will be grouped together as the speech features are used for clustering. Thus, the fine-tuned models will improve the test set of their cluster the most.

Figure 3.2 displays the elbow curve for the selection of k in K-Means clustering. There is no immediately obvious selection, as the inertia seems to go down at a somewhat steady rate, but we select 4 clusters as it is the most feasible option. We then create four

different train-test splits for each cluster and run a comprehensive evaluation of model performance after fine-tuning for all clusters similar to the ones explained in section 3.4.2.

3.4.4 Impact of Language Models

In this section, we investigate if the word error rate reduction achieved by fine-tuning is comparable or better than improvements due to language models on accented speech. Many contemporary ASR systems have language models as part of their pipeline. As we have looked into improvements on the acoustic model performance in previous experiments, this raises the question: Are such improvements redundant if the language models also improve the model performance? Does the baseline Wav2Vec2 model's performance increase significantly enough to make the improvements from our previous experiments negligible?

One of the common language models used in ASR systems is KenLM n-grams due to its indexing and searching for terms more efficiently than traditional language models. This makes it ideal for use in ASR systems [Heafield, 2011].

The `patrickvonplaten/wav2vec2-base-960h-4-gram` model from the HuggingFace website is used for these experiments. This model is identical to the Facebook's baseline Wav2Vec2 960h model, but it is augmented with a Librispeech 4-gram model on top from OpenSLR's website. Since the acoustic and language models are not connected from the pre-training phase, this is a shallow fusion language model [Gülçehre et al., 2015].

The evaluation has been done manually for these, going over samples one by one, calculating each sample's WER, then calculating the average of WERs for all accents as the specific model processor used (Wav2Vec2ProcessorWithLM) cannot be used in HuggingFace's Trainer API, thus making it impossible to use the Trainer API's evaluate function which was used in all other experiments. This might impact the overall results slightly, but we assume the difference would be negligible.

3.4.5 Cross-domain evaluation

In the first four experiments, which evaluated the word error rate performance improvements in the VoxPopuli dataset, we looked into how generalisable these improvements are when the speech data is changed. The VoxPopuli dataset, as mentioned in section 2.3.1, contains speeches from the European Parliament event recordings, likely conducted by experienced orators to an audience and intended as monologues. This is not the usual or common speech that many ASR systems are used for, such as speech recognition in more casual contexts or for human-machine interaction. Therefore, the difference between the speech data in VoxPopuli and many other use cases for ASR systems can be significantly different.

After investigating in-domain improvements on VoxPopuli, we investigate how well the performance improvements compare when the fine-tuned models are presented with

speech data in a different domain and how well the improvements generalise across different domains.

For this experiment, we use the EdAcc [Sanabria et al., 2023] dataset, which is a collection of snippets from phone call conversations between various individuals who already know each other. Therefore, the register is more casual, the tone is more conversational, and the educational and personal backgrounds of individuals are more diverse. The dataset has more information about the speakers' backgrounds as well, but we only take into account the marked raw accent. There are a wide variety of accents, but we select a subset of them for this evaluation. The accents chosen are present in the VoxPopuli dataset (Italian, French, German, etc.), or they are possibly similar to some of the native languages present in the VoxPopuli dataset (Brazilian, Bulgarian, Catalan) or distinctly different from anything seen in the VoxPopuli dataset (Chinese, Egyptian). This provides decent coverage of different accents. Furthermore, there are many short samples in the EdAcc dataset containing less than five words in total. We discard those from our training and evaluation processes. These very short samples disproportionately negatively impact the dataset evaluation. The final practical point is that the EdAcc dataset is sampled at 32 kHz, whereas the model used, Wav2Vec2, can only take audio input sampled at 16 kHz. To make the input formats match, we resample the EdAcc samples using the librosa library in Python.

We ran these cross-domain experiments twice, once with a fine-tuning train set containing 8k words and another using 24k words. The second set of experiments with more data serves multiple purposes. It checks if the fine-tuned model is overfitting to the training data; it provides more insight into how cross-domain learning could work with looser resource constraints (more data) and allows us to check how the model's performance scales with more data across domains.

3.5 Evaluation

The main evaluation metric for models used is word error rate (WER). Although there are different ways of evaluating the performance of ASR systems, such as phoneme error rate or character error rate, word error rate (WER) is the most commonly used one [Yu and Deng, 2016].

Another main characteristic of our evaluation and interpretation of the results throughout the results section is that instead of focusing on raw WER values for accents and experiments, we look into the improvements relative to the baseline WERs. We will report the raw ASR values at some points, too, but if they are not needed to demonstrate a point, I'll put them in the appendix and refer to them there. The main focus of the research here is the improvements and their transferability across accents or domains.

The averages across the dataset are calculated by summing performances or improvements of all accents and dividing by the number of accents. This gives equal weight even though the test set sample sizes may be different across samples. We do this in order to calculate improvements across accents overall instead of the data weighting influencing the reporting of our results (for example, if the performance of French-accented English improves significantly after fine-tuning while other accents do not

Accent	Count	Length	Mean	Median	S.D.	Max.	Mins
Dutch	57	1222	21	17	15	64	7
German	41	1313	32	25	23	99	8
Brazilian	58	1882	32	22	30	134	11
Romanian	94	2127	23	17	18	102	13
Bulgarian	125	3115	25	11	42	239	19
French	83	3379	41	20	51	258	21
European	156	4227	27	18	27	166	26
Chinese	340	6116	18	12	19	171	38
Catalan	242	7804	32	19	36	202	48
Egyptian	282	9252	33	24	30	200	57
East European	525	9518	18	14	15	164	59
Lithuanian	406	9556	24	16	22	207	59
English	571	11722	21	14	19	152	73
Spanish	673	14938	22	15	21	161	93
Italian	458	17318	38	21	48	395	108

Table 3.2: Overview of the EdAcc dataset. Excluding samples containing shorter than 4 words and duplicate values.

show improvement, and if French-accented data is predominant in the test set, this does not affect the performance reporting across all accents.) For accent agnostic setup, we report improvements per cluster, treating each cluster like an accent. In these cases each accent or cluster has varying amounts of data, but mostly enough data to assume that it can generalise further. By treating each accent or cluster equally, we aim to get an overview of accents and improvements to them instead of getting bogged down with differences in the amount of data in each (as most of them have substantial data for testing as well, maybe excepting the lowest few native languages).

Chapter 4

Results

4.1 Fine-tuning on a mix of different accents

For the first experiment, we overview the accent-specific improvements gained from the experiments outlined in section 3.4.1. The raw baseline WERs and the fine-tuned model's WER values are shown in the table 4.2. We present the relative improvements from the baseline in figure 4.1.

The word error rates for different accents using the baseline model range from 0.235 for native English speech to 0.384 for Finnish-accented English. This means that, at best, the ASR system gets over one out of every five words incorrectly, closer to one in every three words for most foreign-accented speech. German- and Dutch-accented English are the best-performing accents apart from native English speech and the only foreign accents with a word error rate below 30

We observed the best improvements when we provided all 8000 words to the model in the training data. The limited amount of data may be a significant constraining factor, thus making every bit of extra data available crucial for the model's learning. This reduces the WER across accents to between 0.198-0.291, with each accent's average accuracy improving by almost 26%. Not every accent improves equally, though; some accents improved by over 30% like Czech, Finnish, or French-accented English, whereas some only 20% or less, like Italian, Spanish, Dutch-accented, or native English speech. This is partly due to the already relatively better performance of the baseline model in some accents (such as Dutch-accented and native English speech) or some accents

Setup	WER Δ
No Validation	**25.98**
75-25 Train-Val Split	22.81
KFolds	24.89

Table 4.1: Average improvements across accents from the baseline by each method (in %)

Accent	Baseline	No Val.	Holdout	KFolds
Czech	0.354	0.236	0.248	0.246
German	0.276	0.209	0.220	0.215
Spanish	0.360	0.291	0.298	0.293
Estonian	0.330	0.235	0.248	0.238
Finnish	0.384	0.261	0.268	0.267
French	0.350	0.239	0.248	0.245
Croatian	0.322	0.249	0.265	0.252
Hungarian	0.319	0.221	0.230	0.228
Italian	0.313	0.260	0.268	0.263
Lithuanian	0.359	0.256	0.270	0.260
Dutch	0.256	0.199	0.210	0.204
Polish	0.358	0.256	0.268	0.263
Romanian	0.319	0.222	0.234	0.227
Slovak	0.318	0.230	0.239	0.234
Slovenian	0.373	0.290	0.297	0.268
English	0.235	0.198	0.204	0.201

Table 4.2: Word Error Rates of experiments on the combined accents database.

possibly being more challenging to improve in general, like with Spanish and Italian, which have high WERs when evaluated with the baseline model. The improvements are not as significant as they are with other accents. We investigate this further by looking into cross-accent improvements for all accents, which may shed further light on the disparities between improvements in different accents.

KFolds experiments were run to check the impact of initialisation sets between four possible iterations of the 75-25 split. By averaging the performance of four training runs, we look at a more representative result accounting for potential differences between different train-validation splits. Average improvement across accents was 24.89%, similar to the experiment run with no validation split. The first 75-25 train-validation split run yielded an average of 22.81% improvements across different accents. This slight difference between the Holdout run and KFolds cross-validation is primarily due to the way averages were calculated. By assuming each accent has similar or equal amount of data, some outliers—namely Slovenian, Croatian, and Lithuanian—can disproportionately affect the overall average. These accents show more pronounced differences between 75-25 train-validation splits and the KFolds results, as shown in figure 4.1. This is most likely due to the fact that these three are the accents with the smallest sample size of all, and due to our prioritisation of samples containing the closest number of words to the average. These accents are likely to include more variation in the training dataset than other accents where the sample size is larger since more data would mean most samples in the training data would contain samples of similar lengths. In contrast, if the sample size is limited, we may run out of average-length samples and resort to using very long or very short samples in the train-validation split.

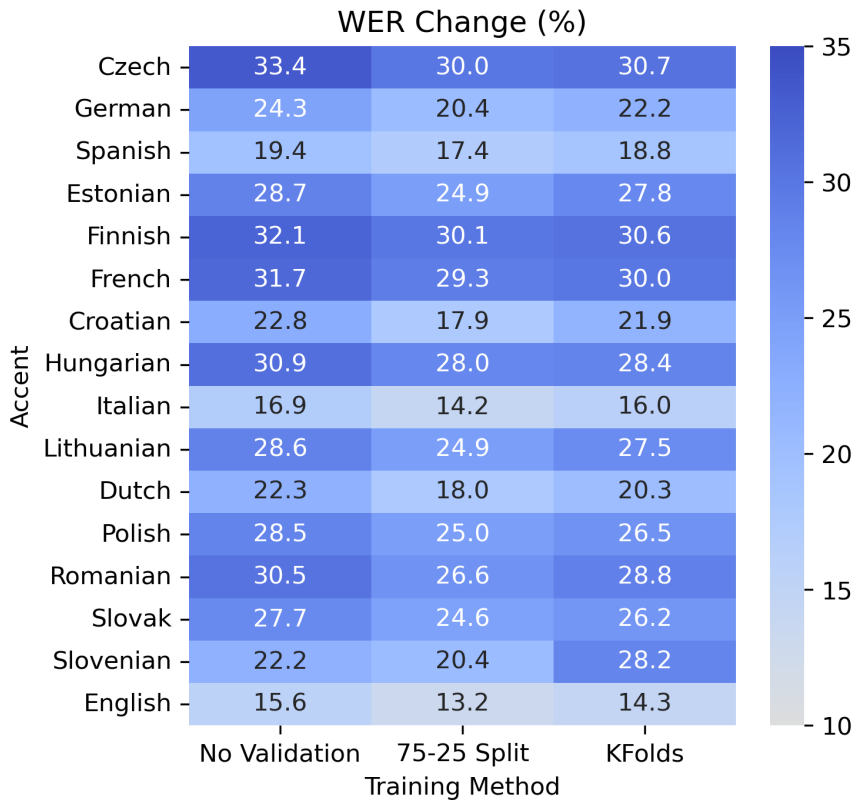

Figure 4.1: Improvements to the test set by accent for each training method.

4.2 Exploring cross-accent learning

Our second set of experiments focused on fine-tuning the baseline model using speech data from single accents. This is a more targeted approach than the first experiment, which contained a mix of different accents. With this, we look into research questions 2 and 3, explained in section 1.2.2. The matrix of difference between fine-tuned performance and the test accent's baseline WER is shown in figure 4.2, and the relative change from the baseline in % points is shown in 4.3. The data with WER values can be found in table A.4 in appendix A.1. For all the tables and figures, training accents are placed on the x axis and test accents on the y axis.

Our hypothesis that each accent's test set is improved the most by the accent-specific trained model holds. When we look at how accent-specific, fine-tuned models improve accents across the board, it is also clear that for most models, the most pronounced improvements are to the test set of the accent they were fine-tuned on. Notable exceptions are German and Dutch-accented English, which increase some other accents more than they improve their respective accents. However, this is due to the already relatively high accuracy of these accents with the baseline model, 27.7% and 25.6%, respectively. When examining the WER values directly (see A.4), the lowest word error rate for each fine-tuned model consistently aligns with the test data of the corresponding accent. Since the same decrease in WER value results in a higher relative change from the baseline for accents with initially higher baseline WERs, this may appear misleading.

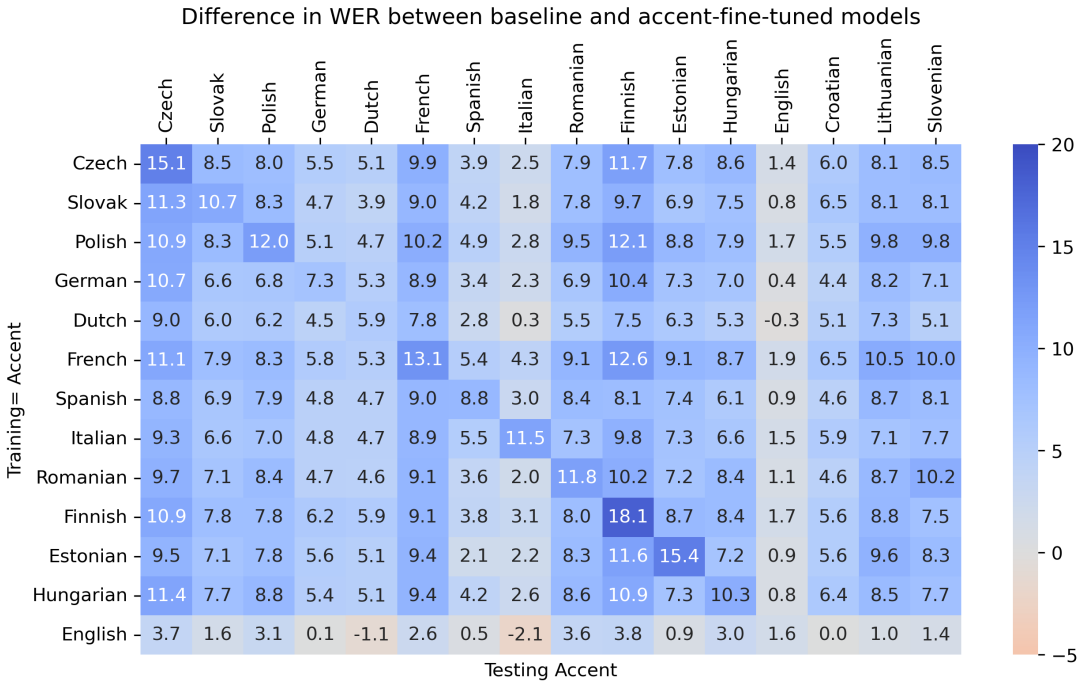

Figure 4.2: Raw WER changes from the baseline values for each accent across fine-tuned models.

To ensure that the fine-tuning improvements are realised because the model learns the accentual differences rather than the dataset-specific jargon, we also fetch a training set from the native-English portion of the VoxPopuli dataset. The native-English fine-tuned model causes minimal improvements across other accents, even causing the performance to degrade in some cases. As we hypothesised in section 3.4.2, fine-tuning improves accents and not necessarily just the model's ability to deal with the domain only. The native-English fine-tuned model resulted in by far the lowest improvements in average, improving the relative performance across accents by only 4%, with the second lowest overall improvements coming from the model fine-tuned with Dutch-accented data at 15.8%.

Another interesting pattern visible in both figures 4.2 and 4.3 is that the improvements from fine-tunings are not always symmetric. Some accents, when provided as training data, improve the overall performance reasonably well, but when other fine-tuned models are evaluated on the accented data, the performance improves noticeably less. Table 4.3 displays an overview of average improvements accented models yield across all accents and average improvements from evaluations with different accented models. While many languages improve other accented speeches, and get improved by models fine-tuned on other accents by similar amounts, there are some outliers. Spanish and Italian for example, improve recognition in all accents by about 19.8% and 21%, respectively. However, other models improve Italian-accented English by only 9% and Spanish-accented English by only 11.3% on average. Some accents show the opposite trend, such as Czech and Finnish. Model fine-tuned with Czech-accented data on average improves performance by 27.4%, but Czech-accented speech is only improved by 22.2% on average, and similarly, the model fine-tuned with Finnish-accented speech

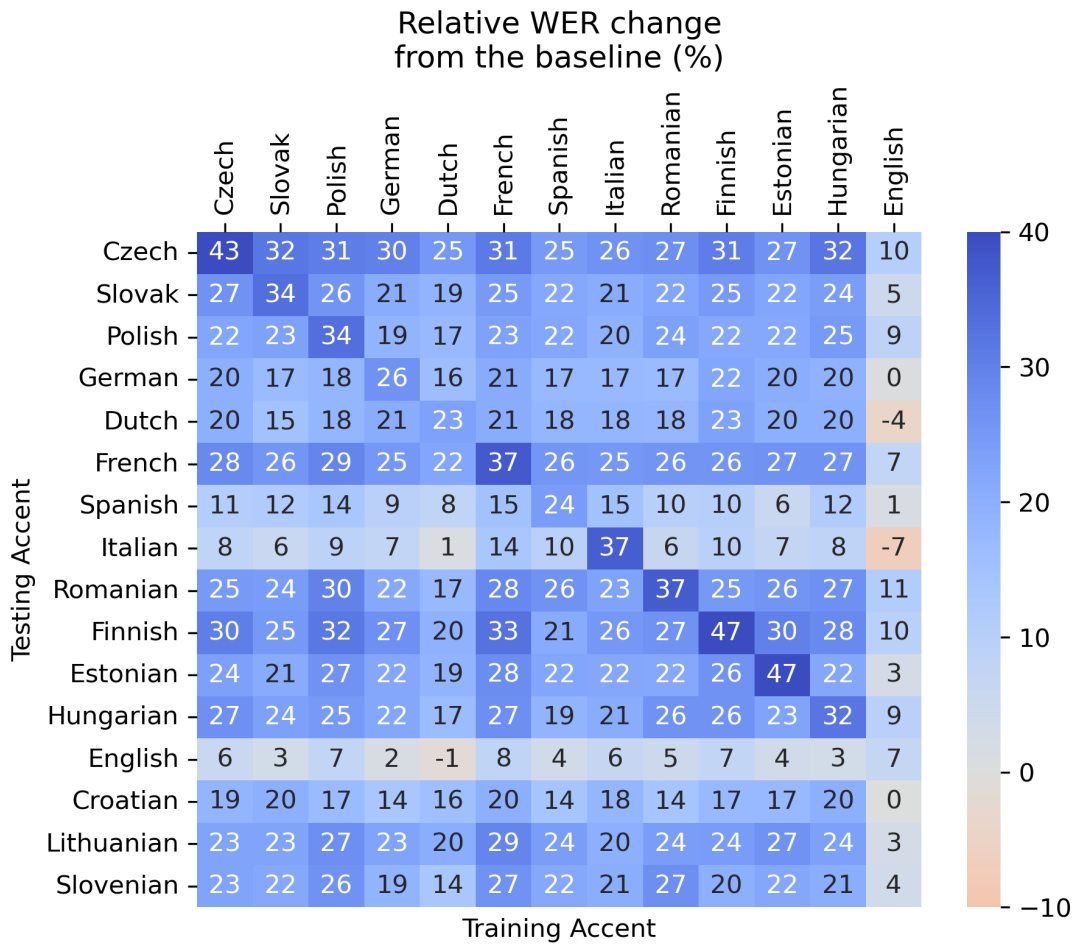

Figure 4.3: Train-test matrix for accent-targeted fine-tuning experiments.

improves performance by 28.5% on average across accents, but other models improve Finnish-accented speech by only 22.7%.

The implications of this disparity could be significant, especially if it reflects a general trend rather than a specific dataset anomaly. This understanding may be valuable during the data collection stages. It is particularly important if certain systems need targeted improvements on specific accents rather than general improvements across multiple accents.

Lastly, we examine the cross-accent improvements within selected subsets of accents. These subsets are grouped based on the language families of the speakers' native accents: Czech, Slovak, and Polish (West Slavic); German and Dutch (Germanic); French, Spanish, Italian, and Romanian (Romance); and Finnish and Estonian (Finnic). Although grouping accents by language families has limitations—such as significant phonetic and linguistic variation even within the same language families—we use this grouping as a heuristic tool for exploring patterns in cross-accent learning. It must be noted that this grouping should not be treated as a definitive linguistic assumption but rather as a means to organise and analyse the data.

Cross-accent learning after fine-tuning overall is a common theme across all runs to

Accent	Training[1]	Testing[2]
French	24.23	25.56
Polish	23.10	21.60
Finnish	22.69	27.35
Czech	22.17	28.53
Estonian	21.65	23.37
Hungarian	21.56	22.91
Italian	21.02	8.95
Romanian	20.81	24.76
Slovak	20.39	22.49
Spanish	19.81	11.34
German	19.32	17.97
Dutch	15.81	17.83
English	4.31	4.71

Table 4.3: Mean improvements to accented sets. [1] How much does the model fine-tuned with this accent improve other accents, [2] On average how much does this accent get improved by models fine-tuned on other accents.

Language Family	Intra-Family Avg	Cross-Family Avg
Slavic	26.87	22.06
Germanic	18.50	16.50
Romance	18.74	22.39
Finnic	28.25	19.66

Table 4.4: Intra-family and cross-family average improvements (in %)

varying extents, often times fine-tuning improving many different accents greatly, both within its language family and other accents from different language families. The patterns are hard to distinguish if we look at accents individually. So, we have collated the average improvements of fine-tuned models within the language family and across other language families. Table 4.4 presents these results concisely. Slavic, Finnic, and Germanic-accented English improve one another within the language family more than they improve accents from other language families, although with Germanic accents, the difference is relatively small. Romance languages seem to perform worse internally, but this is mainly due to the asymmetric nature of Spanish and Italian accents when used as training or testing data. Spanish and Italian accents are useful when provided as fine-tuning data to the model, but they are not affected by other fine-tuned models as much. If we exclude Italian and Spanish from the Romance languages group, the intra-family improvement ratio becomes 27.25%, and the cross-family improvement ratio becomes 19.23%. This shows a pattern of slightly greater intra-language-family improvements, significantly more so for some accents, like Finnic and Slavic language families.

4.3 Accent Agnostic Setup

The MFCC features and Wav2Vec2 embeddings independent of accent labels did not yield strong results. WER results from the baseline model and cluster, along with fine-tuned models for MFCC and Wav2Vec2 clusters, are shown in tables 4.5 and 4.6, respectively. If our hypothesis held, we could expect each fine-tuned model to perform the best results in its respective test data. Ideally, this would cause the diagonal values to be the lowest ones or ones with the greatest amount of relative improvement. This is not the case, however.

When we look into the WER result from the baseline model for clusters, MFCC clusters all have similar WER values. This may mean that the feature extraction and the subsequent grouping resulted in relatively homogenous groups, containing accented speech from across a wide range of accents. This might have caused all the baseline WERs to be similar. This is also a common pattern in the fine-tuned experiments too, where the relative improvement from the baseline model is about 0.07 WER, a relative decrease of around 25%. We look into accent labelled composition of the data in 4.6, it is clear that most of the samples were classified into the 2nd and 3rd clusters, and there is not much clear patterns such as some accents concentrating in certain cluster, which would likely happen if the features were accurately clustered. This distribution means the MFCC features, PCA, and K-Means did not yield a representative enough set of accent-agnostic clusters.

The results from Wav2Vec2 clustering, however, are quite different. They are not as hypothesised as well, but there are more patterns. Looking at the cluster distribution at 4.7, some accents concentrate more on certain clusters. Czech, for example, has 42% of its samples in the 2nd cluster. Italian and Hungarian samples concentrate more on 1st and 2nd clusters; Slovak, Slovenian, and Romanian samples are concentrated in the 2nd cluster; and Finnish, Estonian, and French accented samples are concentrated in the 0th cluster. This is evident in the baseline WERs of clusters as well. Clusters 0 and 1 have

Training Cluster	Baseline	0	1	2	3
Cluster 0	0.315	**0.240**	0.234	0.239	0.236
Cluster 1	0.307	0.234	**0.223**	0.232	0.232
Cluster 2	0.316	0.236	0.230	**0.234**	0.232
Cluster 3	0.317	0.234	0.227	0.233	0**.230**

Table 4.5: MFCC clusters

Training Cluster	Baseline	0	1	2	3
Cluster 0	0.332	**0.231**	0.274	0.206	0.183
Cluster 1	0.382	0.239	**0.275**	0.210	0.196
Cluster 2	0.273	0.242	0.283	**0.210**	0.190
Cluster 3	0.252	0.244	0.290	0.213	**0.190**

Table 4.6: Wav2Vec2 Clusters

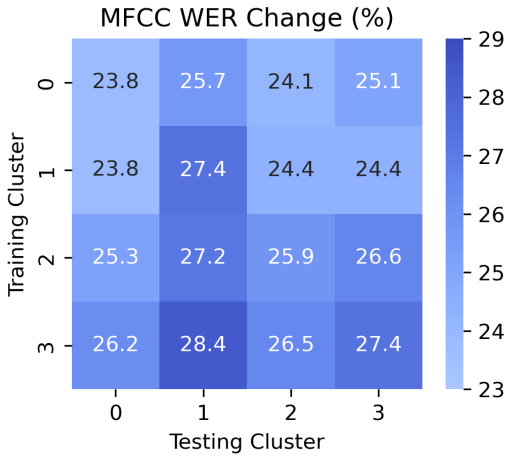

Figure 4.4: Improvements from the baselines to the MFCC clusters.

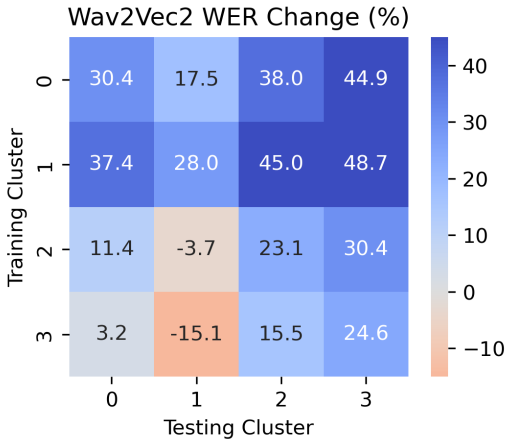

Figure 4.5: Improvements from the baselines to the Wav2Vec2 clusters.

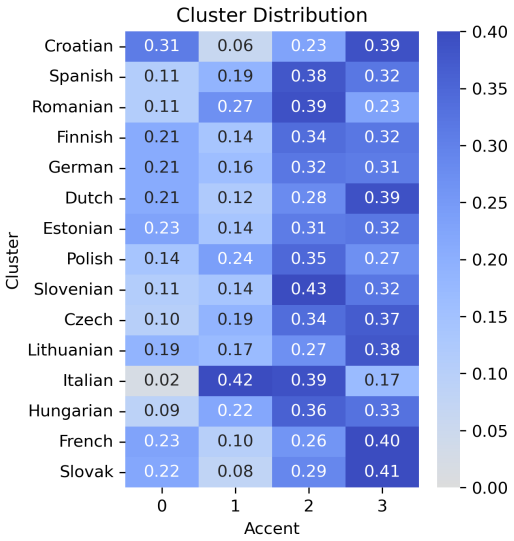

Figure 4.6: Distribution of accented samples across clusters by MFCC features.

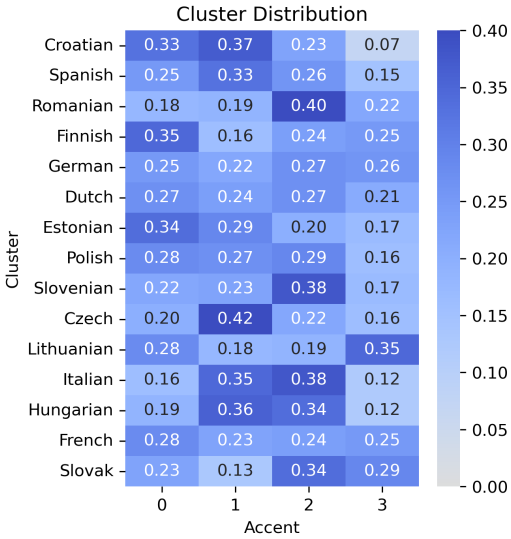

Figure 4.7: Distribution of accented samples across clusters by Wav2Vec2 embeddings.

substantially higher WER values, whereas 2 and 3 have lower than average WER. This may mean the clustering was successful in gathering similar speech samples together.

The improvements from fine-tuning experiments with Wav2Vec2 embedding clusters follow a similar pattern to the MFCC feature clusters. The changes are similar in terms of raw WER reductions from the test set's baseline WER. This may mean that although the accent separation process was possibly successful, the data may still be similar enough that fine-tuned models improve performance across all other clusters similarly, rather than improving their clusters' test set results.

4.3.1 Practical Applicability

The practical motivation for this experiment was to find a middle way between experiments from sections 4.1 and 4.2. Accent-targeted fine-tuning overall improves performance significantly more, but they are computationally intensive and costly if we have to have a model ready for each different accent. The fine-tuning on the combined dataset returns sizeable improvements as well, but not as strongly as the accented ones. The accent-agnostic approach aimed at bridging the gap by abstracting the accent labels. If we only had 4 arbitrary accents that grouped samples together according to speech features, for future use cases, the given input could be classified as well and processed by the fine-tuned model closest to the specific speech's features.

Some issues on this matter, such as outliers, can be problematic both during data processing and when using further data later on. Outlier accent samples can be deemed far away from all other samples; they may be misclassified and transcribed wrongly. Another problem is the inherently data-dependent nature of the PCA and clustering. The clusters may not map well with other datasets or contexts.

4.4 Impact of language models in the pipeline

After implementing the steps described at 3.4.4, our results show that adding a language model increases the accuracy for all accents, for some accents more substantially than others. However, the improvements from the language model do not match the improvements from fine-tuning.

Here, we only examined the impact of a shallow fusion 4-gram KenLM head on top of the Wav2Vec2. Overall, the language model improves the accuracy between 8% and 22.3%. Some accents, like French and Slavic languages, improve more than others. On average, the LM improves the accuracy by 14%, while fine-tuning yields a closer to 27% increase. Table 4.8 shows improvements from the baseline in percentage points. A table of raw WER values in the appendix is also presented in table A.3.

In line with previous findings on language model improvements like Toshniwal et al. [2018], Kumar and Niranjan [2024], our model shows that although general language models are beneficial, a targeted labelled fine-tuning data, even if it is a small one (of 8000 words in our experiment, which corresponds to roughly an hour of speech), can be significantly more impactful.

4.5 Cross-domain generalisation of learning

Our last batch of experiments delve into the question of generalisation of the fine-tunings between domains. So far, the experiments and evaluation of improvements were all run with data from the VoxPopuli dataset. And as mentioned previously, although this dataset has a wide variety of European English accents, the topic diversity is limited, the context of the speech samples are all monologues in the European Parliament events, and the register of speech is relatively more formal than many other use cases where ASR may be needed. To test the performance of accented speech in a different domain,

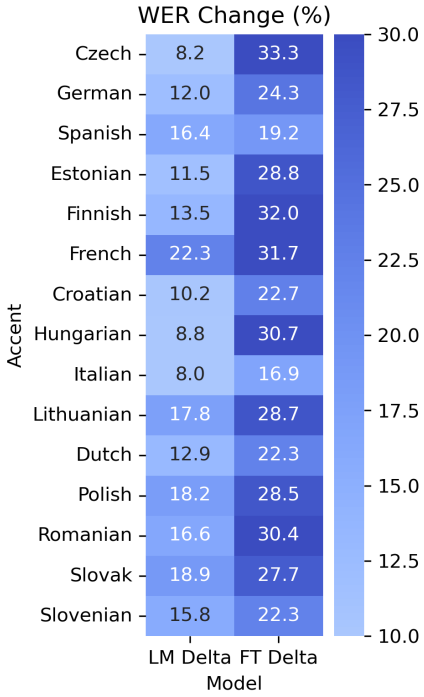

Figure 4.8: Comparison of improvements gained from using a language model on the acoustic model vs fine-tuning.

we use the EdAcc [Sanabria et al., 2023] dataset, which contains speech from phone conversations between people who know each other. This provides a highly contrasting dataset with a more casual register and a conversational tone. EdAcc overall has a much higher average WER for most of the accents with the baseline Wav2Vec2 model, with many accents having a WER of around 50% or more. This is also quite a contrast and something to be aware of when looking at results here. Since we report performance changes relative to the baseline values, the same percentage change in VoxPopuli and EdAcc can amount to different raw WER changes.

Our evaluation of the cross-domain performance of the fine-tuned models indicates that there are little or negligible improvements to accented speech in different domains. The results from fine-tuning the baseline model with VoxPopuli data is shown in figure 4.11, including performance increases to the in-domain VoxPopuli dataset as well as the cross-domain EdAcc dataset. In-domain improvements to the datasets are at 27% for VoxPopuli, and 20% for EdAcc datasets when fine-tuned using 8k words.

The default training size used in all the previous experiments interestingly caused the accuracy to degrade in some accents in the EdAcc dataset, like German and Egyptian accents; the most improved accent was Brazilian-accented English, which saw a 12.6% increase in its accuracy with the default setup.

After seeing that the amount of training data that caused significant improvements to the in-domain data failed to show significant improvements in cross-domain samples, we decided fine-tune the models again, this time using three times more data, 24k words in total, to see how much this would improve the performance across accents in the in-

Accent	Baseline	8k Words	24k Words
Italian	0.391	0.377	0.345
French	0.533	0.518	0.505
German	0.303	0.330	0.300
Brazilian	0.496	0.434	0.405
Romanian	0.633	0.629	0.618
Egyptian	0.384	0.396	0.366
Chinese	0.542	0.534	0.525
Lithuanian	0.533	0.501	0.475
Spanish	0.516	0.516	0.496
Bulgarian	0.567	0.538	0.511
European	0.492	0.483	0.467
Catalan	0.607	0.573	0.541
Dutch	0.537	0.502	0.475
East European	0.470	0.459	0.437

Table 4.7: WERs on EdAcc dataset.

and cross-domain test sets. Overall, more data helps in cross-domain generalisation, but the improvement stays relatively low compared to in-domain gains. In terms of improvements to each accent, the accents included in the training data show relatively high improvements in the cross-domain set, too, although there are exceptions like Romanian, French, and Spanish, which show minimal improvements in the cross-domain set.

Lastly, we swap the train and test sets, looking into the transferability of learning from a more casual and conversational domain to a more formal and monologue-based domain. The results of this experiment are shown in figure 4.14. Similarly to before, we see significant accuracy improvements for the in-domain training set. The highest improvement is seen in Brazilian-accented English again, although this may potentially be oversaturated as most of the data for Brazilian-accented English is in the training data. Thus, the test data for Brazilian-accented English is smaller than many other accents' test sets. The cross-domain evaluation of the model fine-tuned on EdAcc data shows only minimal overall improvements. However, it performs slightly better than the model fine-tuned on VoxPopuli when tested on EdAcc. This suggests that models trained on more casual and permissive domains may generalise better to stricter domains. In contrast, fine-tuning on more specific domains can sometimes negatively impact performance across accents in different domains.

This is in line with the literature, fine-tuning typically does not transfer across domains and may lead to performance degradation. We see minimal overall gains between domains, with some accuracy loss in accented speech, especially when fine-tuning data is more domain-specific [Paraskevopoulos et al., 2023, Zhou et al., 2023].

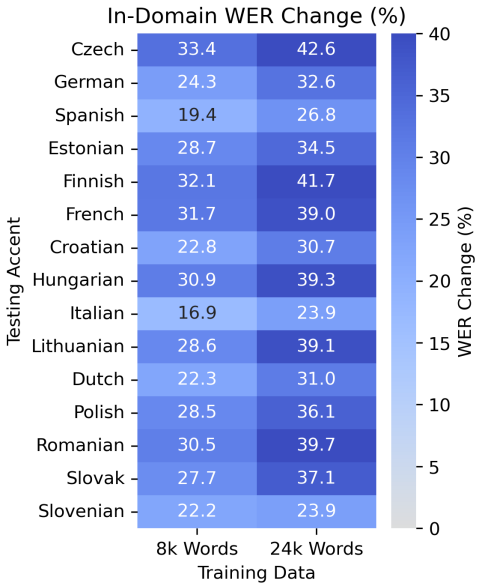

Figure 4.9: Fine-tuned on combined VoxPopuli training set, evaluated on VoxPopuli test set.

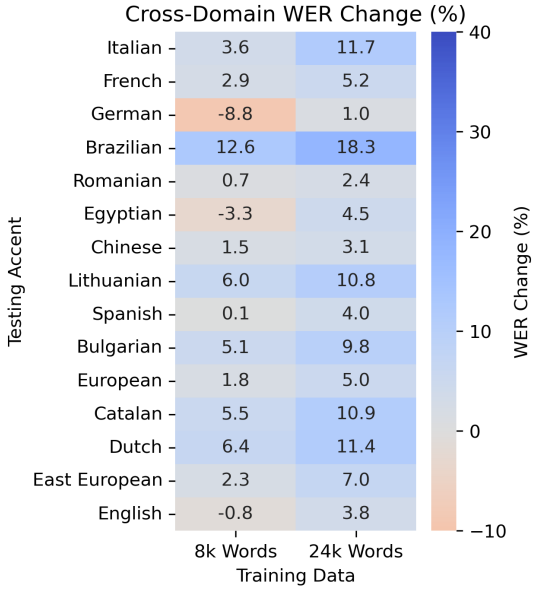

Figure 4.10: Fine-tuned on combined VoxPopuli training set, evaluated on the filtered EdAcc test set.

Figure 4.11: Improvements achieved by the model fine-tuned on data from combined VoxPopuli training set. Figures are percentage increases from the baseline WER values.

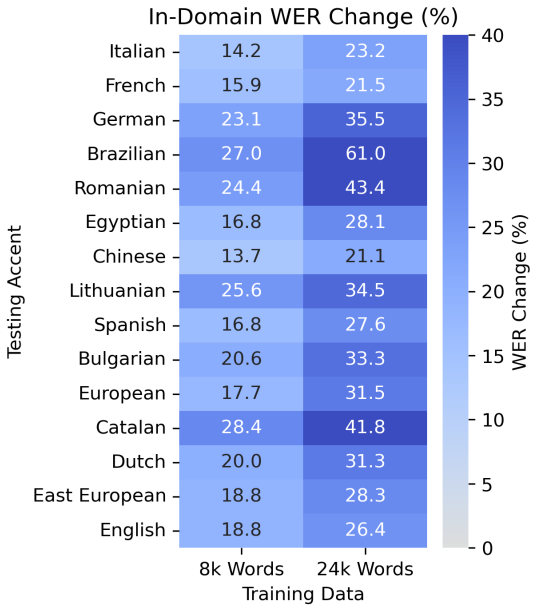

Figure 4.12: Fine-tuned on EdAcc training set, evaluated on EdAcc test set.

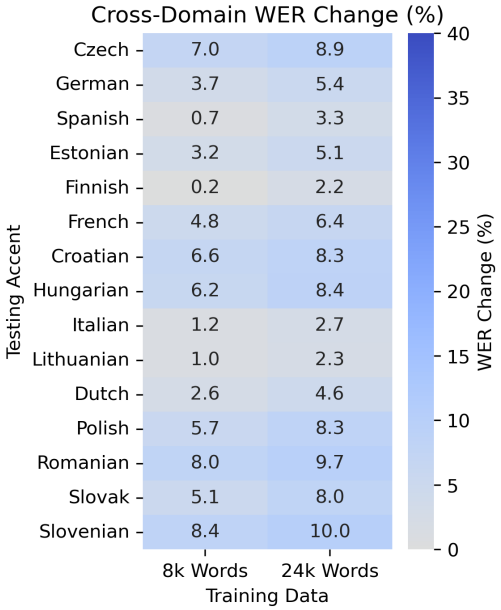

Figure 4.13: Fine-tuned on EdAcc training set, evaluated on VoxPopuli.

Figure 4.14: Improvements achieved by the model fine-tuned on data from combined EdAcc training set. Figures are percentage increases from the baseline WER values.

Chapter 5

Discussions

5.1 Limitations

Although the results show a significant potential for improvement in accented speech recognition and cross-accent transferability of learning, there are still numerous limitations to this study. First and foremost, we only focus on predominantly European foreign-accented English in our datasets. Using speech from people who learn a dominant variant of English as a second language. This study does not examine any native English speakers with non-standard accents. These accents can be subtle and easy to understand for many, like Canadian or some British or American accents that are similar to the "standardised" ones, but there are also certain regional accents that are traditionally seen as much harder due to strong divergence from mainstream English pronunciation rules, regional lexicon and phrases, and unusual grammatical formations such as Glaswegian, Irish, or Jamaican accents [Alfonso Durbán, 2018]. Additionally, the dataset we used throughout the paper, *VoxPopuli*, contains speeches from a limited set of speakers, many possibly coming from highly educated backgrounds due to their occupation, position, and nature of the recordings. Although this does not mean they do not have accents or inherently have better accents, it is a reasonable concern that their educational background may not be as representative of the general population.

Another limitation of the data was that almost no samples were labelled according to the genders or age groups of the speakers. Although ASR systems do show biases against certain demographics [Jahan et al., 2025, Fuckner et al., 2023], these biases were not addressed or included in the evaluation of our experiments.

Lastly, the context and domain of the dataset used is a niche one. European Parliament speeches are all delivered in the format of a monologue. Many of the samples will likely contain words, phrases, and sentences about political or policy debates. This domain-specific nature and formal monologue register of the speeches can also be a limiting factor in the applicability of the results to other domains, contexts, or registers [Zhou et al., 2023]. We have attempted to address this and explore cross-domain transfer of the fine-tuning with the *EdAcc* dataset, and it has shown this limitation as a potential factor negatively impacting the cross-domain generalisation.

5.2 Concerns

There are also numerous concerns this study and its results may cause. Should there be a generalisable trend with the transfer of learning between similar accents, this can discourage further research into low-resource accents if there is a closer and easier-to-gather data accent that is more available to the researchers. Potential propagation of exclusion of minorities from the research focus by leveraging more common high-resource accents may be detrimental to the speakers of the accent unintentionally, causing the accent to receive less attention from the research community. This can also be disturbing due to historical context, For example, it might make sense to use Russian-accented English to improve Ukrainian-accented English due to the fact that they are both eastern Slavic languages with similar lexicon and linguistic qualities, but the socio-political context could make this use case quite repulsive and unacceptable by the speakers of these languages.

5.3 Future Work

Our experiments demonstrate considerable improvements to in-domain accented English speech data through fine-tuning using a pre-trained ASR system. Some improvements and additions could still be introduced into these experiments to explore different ways of optimising ASR systems' performance across accents and domains. Several papers suggest the introduction of accented data or foreign-language data [Shankar et al., 2018, Kumar et al., 2023] in the pre-training stage. This might be a beneficial suggestion to further explore, especially more so as it can highlight accent-recognition challenges from an earlier design stage for ASR practitioners.

5.3.1 Accent-Agnostic

Accent-agnostic experiments suffered from inconclusive results in this paper. Although the improvements achieved in average are similar to the combined and accent-specific improvements, the speech-cluster methods were too simplistic and possibly due to the severe data loss during principal component analysis. More sophisticated techniques, such as the AccentFusion framework [Gu et al., 2024] can be good starting points for future work on leveraging accent similarity for cross-accent learning.

Another way could be to create an accent classifier with self-supervision and, again, abstract clusters. This classifier could be a multilayer neural network, using the MFCC features, Wav2Vec2 or Speech2Vec embeddings as input, and categorising speech into a group, which could then be further used to fine-tune and improve these accented speeches based on similar samples.

For accent agnostic experiments, more could be done to investigate the concept, with potentially stronger ways of clustering accents together that take in more information than what was tested. Another potential way of using this system could be to dynamically allocate more data to the clusters that have worse initial WER to improve their fine-tuning performance more than clusters that have already ok (relatively) performing WER.

5.3.2 Language Model

The language model, incorporated into the neural and transformer-based ASR systems through deep, shallow, or cold fusion, has shown improvements to the model accuracy in our experiments and in previous work [Gülçehre et al., 2015, Sriram et al., 2017, Kumar and Niranjan, 2024]. Furthermore, we could look into utilising the power of neural language models or transformer-based architectures alongside fine-tuning these models with labelled speech data similarly to Sullivan et al. [2022], Liu et al. [2024b], Arisoy et al. [2015], Toshniwal et al. [2018]. This could improve our models' accuracy beyond the current improvements reported here. While some uses of the large language model have reported increased accuracy, especially earlier models such as BERT [Chiu and Chen, 2021], some other research has reported increased WER [Min and Wang, 2023], this is another ongoing research area, and could benefit from further work.

5.3.3 Cross-Domain Adaptation

Although some preliminary work on examining the generalisation of this training into a different domain has been done in the section 4.5, its scale and extent are quite limited. We also do not implement or offer ways of tackling this problem. There are numerous studies on how domain generalisation in ASR and in accented speech can be utilised to address such problems [Zhou et al., 2022, Paraskevopoulos et al., 2023, Zhou et al., 2023]

Various previous work influenced by meta-learning [Finn et al., 2017, Li et al., 2018, Zhou et al., 2023] have shown promising results in robust, domain-generalisable, and efficient methods. These can be beneficial in addressing both domain-adaptation in terms of context, but also accent.

5.4 Final remarks

In conclusion, our experiments demonstrate significant improvements in recognising accented English speech within the same domain of speech. Even with a limited dataset of approximately one hour of speech, we achieved substantial gains in performance. We also explored the trade-offs between fine-tuning the model with a mix of accents and using a more focused, accent-specific dataset. Our findings indicate that the most significant improvements in accuracy for some accents occur with accent-targeted training; however, utilising a combined dataset can also be an effective strategy for enhancing performance across a variety of accents.

Additionally, we show that fine-tuning has a more pronounced impact on the model's accuracy compared to incorporating language models in the processing pipeline. Despite these strengths, our experiments also highlight a critical issue with domain adaptation through fine-tuning: models struggle to generalise well across different domains.

Bibliography

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. Accentdb: A database of non-native english accents to assist neural speech recognition. *arXiv preprint arXiv:2005.07973*, 2020.

Mireia Alfonso Durbán. Interpreting accents: An analysis of the cognitive process of interpreting the scottish accent taking a phonological approach. 2018.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

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

Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen. Bidirectional recurrent neural network language models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5421–5425. IEEE, 2015.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving,

multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, et al. Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*, 2017.

Shih-Hsuan Chiu and Berlin Chen. Innovative bert-based reranking language models for speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–271. IEEE, 2021.

Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*, 2018.

Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

David Crystal. *English as a global language*. Cambridge university press, 2003.

Amit Das, Jinyu Li, Rui Zhao, and Yifan Gong. Advancing connectionist temporal classification with attention modeling. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 4769–4773. IEEE, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151. IEEE, 2023.

Hayato Futami, Hirofumi Inaguma, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Distilling the knowledge of bert for sequence-to-sequence asr. *arXiv preprint arXiv:2008.03822*, 2020.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Hongjie Gu, Gang Sun, Ran Shen, Yifan Wang, Weihao Jiang, and Junjie Huang. Exploring accent similarity for cross-accented speech recognition. In *Proceedings of the 2024 8th International Conference on Control Engineering and Artificial Intelligence*, pages 232–237, 2024.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535, 2015. URL `http://arxiv.org/abs/1503.03535`.

Imad Qasim Habeeb, Hanan Najm Abdulkhudhur, and Zeyad Qasim Al-Zaydi. Three n-grams based language model for auto-correction of speech recognition errors. In *International Conference on New Trends in Information and Communications Technology Applications*, pages 131–143. Springer, 2021.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197, 2011.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 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 20*, pages 198–208. Springer, 2018.

Louis Hickman, Markus Langer, Rachel M Saef, and Louis Tay. Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*, 2024.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*, 2021.

Teemu Hirsimaki, Janne Pylkkonen, and Mikko Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):724–732, 2009.

Sho Inoue, Shuai Wang, Wanxing Wang, Pengcheng Zhu, Mengxiao Bi, and Haizhou Li. Macst: Multi-accent speech synthesis via text transliteration for accent conversion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Kazuki Irie, Zoltán Tüske, Tamer Alkhouli, Ralf Schlüter, Hermann Ney, et al. Lstm, gru, highway and a bit of attention: An empirical overview for language modeling in speech recognition. In *Interspeech*, pages 3519–3523, 2016.

Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. Unveiling performance bias in asr systems: A study on gender, age, accent, and more. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1998.

Biing-Hwang Juang and Lawrence R Rabiner. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1(67):1, 2005.

Jodi Kearns. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8, 2014.

John Kominek and Alan W Black. The cmu arctic speech databases. In *SSW*, pages 223–224, 2004.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE, 2020.

Mehul Kumar, Jiyeon Kim, Dhananjaya Gowda, Abhinav Garg, and Chanwoo Kim. Self-supervised accent learning for under-resourced accents using native language data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

MJ Dileep Kumar and Prabha Niranjan. Development of a language model to enhance the performance of hindi automatic speech recognition. In *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 836–842. IEEE, 2024.

Max WY Lam, Xie Chen, Shoukang Hu, Jianwei Yu, Xunying Liu, and Helen Meng. Gaussian process lstm recurrent neural network language models for speech recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7235–7239. IEEE, 2019.

T. Lander. Cslu: Foreign accented english release 1.2. URL `https://catalog.ldc.upenn.edu/LDC2007S08`.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*, 2019.

Zoey Liu, Nitin Venkateswaran, Eric Le Ferrand, and Emily Prud'hommeaux. How important is a language model for low-resource ASR? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 206–213, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.13. URL `https://aclanthology.org/2024.findings-acl.13/`.

Zoey Liu, Nitin Venkateswaran, Éric Le Ferrand, and Emily Prud'hommeaux. How important is a language model for low-resource asr? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 206–213, 2024b.

Jian Luo, Jianzong Wang, Ning Cheng, Edward Xiao, Jing Xiao, Georg Kucsko, Patrick O'Neill, Jagadeesh Balam, Slyne Deng, Adriana Flores, et al. Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80: 9411–9457, 2021.

Nina Markl and Catherine Lai. Everyone has an accent. In *Interspeech 2023*, pages 4424–4427. ISCA, 2023.

Nina Markl and Stephen Joseph McNulty. Language technology practitioners as language managers: arbitrating data bias and predictive bias in asr. *arXiv preprint arXiv:2202.12603*, 2022.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

Zeping Min and Jinbo Wang. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. In *International Conference on Neural Information Processing*, pages 69–84. Springer, 2023.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

Mikel K Ngueajio and Gloria Washington. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International conference on human-computer interaction*, pages 421–440. Springer, 2022.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Georgios Paraskevopoulos, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for modern greek. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 286–299, 2023.

Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

Nicole Hodges Persley. An innovative idea: a review of the international dialects of english archive. *English Today*, 29(3):63–64, 2013. doi: 10.1017/S0266078413000229.

Aleksander Pohl and Bartosz Ziółko. Using part of speech n-grams for improving automatic speech recognition of polish. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 492–504. Springer, 2013.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Carlo Scagliola. Language models and search algorithms for real-time speech recognition. *International Journal of Man-Machine Studies*, 22(5):523–547, 1985.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training, 2018. URL `https://arxiv.org/abs/1804.10745`.

Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922. IEEE, 2021.

Yuvika Singh, Anban Pillay, and Edgar Jembere. Features of speech audio for accent recognition. In *2020 International conference on artificial intelligence, big data, computing and data communication systems (icABCD)*, pages 1–6. IEEE, 2020.

Martina Slámová. *Germanic and Slavic Accents of English*. PhD thesis, Masaryk University, Faculty of Arts, 2018.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models. *CoRR*, abs/1708.06426, 2017. URL `http://arxiv.org/abs/1708.06426`.

Peter Sullivan, Toshiko Shibano, and Muhammad Abdul-Mageed. Improving automatic speech recognition for non-native english with transfer learning and language model decoding. In *Analysis and application of natural language and speech processing*, pages 21–44. Springer, 2022.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Steve Tauroza and Desmond Allison. Speech rates in british english. *Applied linguistics*, 11(1):90–105, 1990.

Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pages 369–375. IEEE, 2018.

Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE, 2021.

Joost Van Doremalen, Lou Boves, Jozef Colpaert, Catia Cucchiarini, and Helmer Strik. Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning*, 29(4):833–851, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.acl-long.80`.

Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv:2406.14875*, 2024.

Steven H Weinberger and Stephen A Kunath. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*, pages 265–281. Brill, 2011.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Dong Yu and Lin Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. In *Interspeech*, pages 2783–2787, 2018.

Jiaming Zhou, Shiwan Zhao, Ning Jiang, Guoqing Zhao, and Yong Qin. Madi: Inter-domain matching and intra-domain discrimination for cross-domain speech recog-

nition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *arXiv preprint arXiv:2305.18283*, 2023.

Appendix A

First Appendix

A.1 Extras

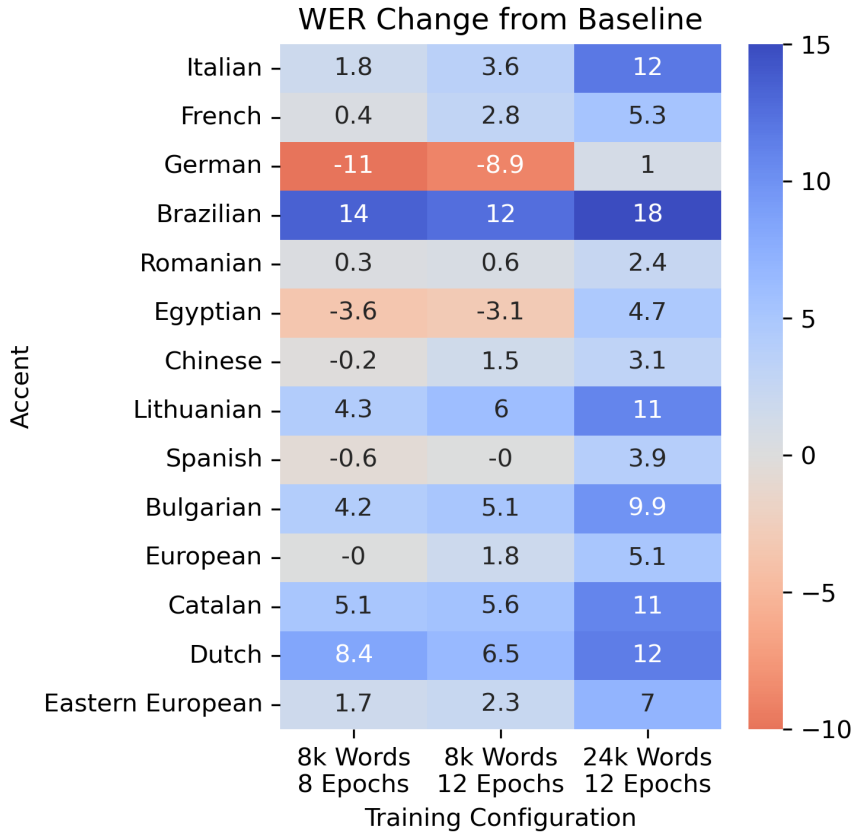

Figure A.1: TODO Elaborate more

Cluster	0		1		2		3	
Split	Train	Test	Train	Test	Train	Test	Train	Test
Czech	18	76	50	133	38	291	46	311
German	48	178	23	149	37	302	26	300
Spanish	4	42	16	64	15	141	9	122
Estonian	13	52	8	32	7	80	13	78
Finnish	28	92	16	65	22	173	20	166
French	31	140	11	61	28	165	37	257
Croatian	10	30	3	5	4	25	3	47
Hungarian	10	47	26	119	21	222	26	198
Italian	2	6	36	111	20	116	8	53
Lithuanian	10	23	9	21	8	39	10	57
Dutch	60	179	20	117	25	299	43	396
Polish	25	94	38	165	31	267	18	208
Romanian	5	49	22	105	19	169	11	98
Slovak	27	64	5	26	14	105	19	146
Slovenian	2	5	6	3	3	25	0	21

Table A.1: Train-test data splits by accent labels for MFCC-based clusters.

Cluster	0		1		2		3	
Split	Train	Test	Train	Test	Train	Test	Train	Test
Czech	37	153	72	329	30	185	13	144
German	34	230	21	214	41	249	38	236
Spanish	12	92	8	130	20	88	4	59
Estonian	11	85	12	70	10	46	8	41
Finnish	20	184	19	73	26	116	21	123
French	34	170	24	145	34	141	15	167
Croatian	4	38	6	41	8	21	2	7
Hungarian	22	102	25	218	29	196	7	70
Italian	12	45	20	102	30	102	4	37
Lithuanian	10	40	8	24	4	29	15	47
Dutch	42	268	34	244	47	265	25	214
Polish	36	203	26	203	31	213	19	115
Romanian	6	82	12	81	29	161	10	97
Slovak	12	83	10	44	17	121	26	93
Slovenian	1	13	2	13	6	19	2	9

Table A.2: Train-test data splits by accent labels for Wav2Vec2-based clusters

Accent	Without LM	With LM	Fine-tuned
Czech	0.354	0.325	0.236
German	0.276	0.243	0.209
Spanish	0.360	0.301	0.291
Estonian	0.330	0.292	0.235
Finnish	0.384	0.332	0.261
French	0.350	0.272	0.239
Croatian	0.322	0.289	0.249
Hungarian	0.319	0.291	0.221
Italian	0.313	0.288	0.260
Lithuanian	0.359	0.295	0.256
Dutch	0.256	0.223	0.199
Polish	0.358	0.293	0.256
Romanian	0.319	0.266	0.222
Slovak	0.318	0.258	0.230
Slovenian	0.373	0.314	0.290

Table A.3: Impact of having a language model on the ASR model.

	Baseline	Czech	German	Spanish	Estonian	Finnish	French	Hungarian	Italian	Dutch	Polish	Romanian	Slovak	English
Czech	0.354	0.203	0.247	0.266	0.259	0.245	0.243	0.240	0.261	0.264	0.245	0.257	0.241	0.317
German	0.276	0.221	0.203	0.228	0.220	0.214	0.218	0.222	0.228	0.231	0.225	0.229	0.229	0.275
Spanish	0.360	0.321	0.326	0.272	0.339	0.322	0.306	0.318	0.305	0.332	0.311	0.324	0.318	0.355
Estonian	0.330	0.252	0.257	0.256	0.176	0.243	0.239	0.257	0.257	0.267	0.242	0.258	0.261	0.321
Finnish	0.384	0.267	0.280	0.303	0.268	0.203	0.258	0.275	0.286	0.309	0.263	0.282	0.287	0.346
French	0.350	0.251	0.261	0.260	0.256	0.259	0.219	0.256	0.261	0.272	0.248	0.259	0.260	0.324
Croatian	0.322	0.262	0.278	0.276	0.266	0.266	0.257	0.258	0.263	0.271	0.267	0.276	0.257	0.322
Hungarian	0.319	0.233	0.249	0.258	0.247	0.235	0.232	0.216	0.253	0.266	0.240	0.235	0.244	0.289
Italian	0.313	0.288	0.290	0.283	0.291	0.282	0.270	0.287	0.198	0.310	0.285	0.293	0.295	0.334
Lithuanian	0.359	0.278	0.277	0.272	0.263	0.271	0.254	0.274	0.288	0.286	0.261	0.272	0.278	0.349
Dutch	0.256	0.205	0.203	0.209	0.205	0.197	0.203	0.205	0.209	0.197	0.209	0.210	0.217	0.267
Polish	0.358	0.278	0.290	0.279	0.280	0.280	0.275	0.270	0.288	0.296	0.238	0.274	0.275	0.327
Romanian	0.319	0.240	0.250	0.235	0.236	0.239	0.228	0.233	0.246	0.264	0.224	0.201	0.241	0.283
Slovak	0.318	0.233	0.252	0.249	0.247	0.240	0.239	0.241	0.252	0.258	0.235	0.247	0.211	0.302
Slovenian	0.373	0.288	0.302	0.292	0.290	0.298	0.273	0.296	0.296	0.322	0.275	0.271	0.292	0.359
English	0.235	0.221	0.231	0.226	0.226	0.218	0.216	0.227	0.220	0.238	0.218	0.224	0.227	0.219

Table A.4: Accented tests WER values

Accent	Baseline	8k8e	8k12e	24k12e
Italian	0.434	0.476	0.46 2	0.437
French	0.601	0.591	0.592	0.581
German	0.403	0.425	0.419	0.383
Brazilian	0.536	0.484	0.481	0.467
Romanian	0.735	0.739	0.738	0.715
Egyptian	0.421	0.430	0.419	0.402
Chinese	0.621	0.670	0.658	0.631
Lithuanian	0.559	0.552	0.539	0.509
Spanish	0.568	0.581	0.574	0.556
Catalan	0.641	0.626	0.614	0.566
Bulgarian	0.664	0.655	0.638	0.606
European	0.562	0.563	0.560	0.541
Dutch	0.629	0.608	0.590	0.564
Eastern European	0.527	0.544	0.530	0.506

Table A.5: WERs on EdAcc dataset with various configurations of training data and length (8 or 24 thousand words, and 8 or 12 epochs relatively.)