university of groningen

campus fryslân

Enhancing English Dysarthric Speech Recognition with Age-Matched Healthy Speech: A Fine-Tuning Approach Using wav2vec 2.0

Cantao Su

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Enhancing English Dysarthric Speech Recognition with Age-Matched Healthy Speech: A Fine-Tuning Approach Using wav2vec 2.0

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen, under the supervision of
Asst. Prof. Dr. Vass Verkhodanova (Voice Technology, University of Groningen)
with the second reader being
TBD (Voice Technology, University of Groningen)

Cantao Su (S4802829)

June 30, 2024

Acknowledgements

I would like to express my gratitude to my first supervisor Vass Verkhodanova and my second reader Shekhar Nayak. Brainstorming with them was instrumental in refining my research questions, and their valuable feedback during the later stages of this thesis significantly improved the quality of my work. Vass' detailed suggestions on my thesis revisions were especially invaluable; her thorough and insightful comments helped to enhance the clarity and depth of my arguments. The extensive effort she put into reviewing my drafts and providing constructive criticism played a crucial role in shaping the final version of this thesis. I also extend my thanks to Tan Phat Do for his technical supports and advice, which made the training of my models much easier.

A special mention goes to Duy Khanh Le, the creator of the GitHub repository 'ASR-Wav2vec-Finetune', Leo Yang, the co-founder of the S3PRL toolkit, and Sanyuan Chen, the original author of the WavLM model. Their work provided substantial technical guidance, and their willingness to help with my questions was invaluable.

Most importantly, I would like to acknowledge my own unwavering dedication and perseverance throughout these past three months of model training and thesis writing. Many nights were spent working until the early hours of the morning, driven by a relentless effort and a commitment to overcoming challenges. This determination was instrumental in bringing this thesis to fruition, and I take immense pride in the personal growth and academic achievements realised during this transformative process.

Lastly, I would like to extend my heartfelt gratitude to all the professors and peers I have had the pleasure of meeting throughout this master's programme. Their collective wisdom, encouragement, and camaraderie have greatly enriched my academic journey and contributed significantly to my success.

Abstract

Automatic Speech Recognition (ASR) has made significant advancements since its advent, particularly in recent years. However, ASR for dysarthric speech remains a substantial challenge due to its high variability and the limited labelled data for training. This thesis focuses on the fine-tuning phase of the wav2vec 2.0 model, which is pre-trained on large-scale English datasets, aiming to improve the recognition accuracy of dysarthric speech. Specifically, this study investigates the impact of incorporating age-matched healthy speech during the fine-tuning process.

Utilising the TORGO dataset, which includes dysarthric speech from speakers with cerebral palsy (CP) and amyotrophic lateral sclerosis (ALS) alongside non-dysarthric controls, this thesis evaluates the performance of ASR models fine-tuned with and without age-matched healthy speech. The methodology involves comparing models fine-tuned with dysarthric speech alone, dysarthric speech combined with age-matched healthy speech, and dysarthric speech combined with age-unmatched healthy speech.

In addition to speaker-independent settings, this study also expands to speaker-dependent scenarios by fine-tuning and validating models on speech data from individual dysarthric speakers with varying levels of intelligibility. This approach provides a comprehensive evaluation of the models' performance across different severity levels of dysarthria.

The results of this research provide practical insights into the effectiveness of incorporating age-matched healthy speech data in training robust ASR models for dysarthric speech. By leveraging the strengths of wav2vec 2.0 and utilising age-matched data, this work aims to contribute to the development of more accurate and reliable ASR systems for individuals with speech impairments. Ultimately, this research seeks to improve the accessibility of voice technology and communication for affected populations.

Contents

1 Introduction

The advent of digital communications marked the beginning of a new era, where obstacles to information access and human interaction are gradually being broken down. With the development of speech technology, people are increasingly relying on voice interaction to accomplish everyday tasks[1]. However, the benefits of this technological revolution has not expanded to all members of society equally.

For some groups, such as linguistic minorities and people with special speech needs, existing speech recognition systems may not be as effective. Out of the 6,912 existing languages in the world[2], only 146 are supported by Google Cloud Speech-to-Text V2[1]. A language becomes a low-resource not always due to the low number of speakers, but sometimes due to the lack of commercial profitability. Even with a large number of speakers, if the country has a low level of economic development, the technology for that language may remain underexplored. For example, the languages of some African and Asian countries with large populations but relatively underdeveloped economies are considered understudied[3]. The issue of limited research is a concern not just for minoritized languages, but also for minority communities. Due to their small populations and the limited availability of data and commercial incentives, these groups often find their requirements for speech technology being overlooked.

People with dysarthria are one of these groups. Dysarthria is a motor speech disorder caused by neurological damage to the motor part of the motor-speech system, which can severely affect speech intelligibility and, consequently, the ability to interact with voice-activated technology[4]. In recent years, there has been an increase in publicly available databases of dysarthric speech, such as TORGO[2], UASpeech[3], and Nemours[5]. Alongside, advancements in data augmentation techniques[6] and sophisticated modeling and training strategies have contributed to some progress in Automatic Speech Recognition (ASR) research on dysarthric speech[7, 8, 9, 10, 11].

Among these studies, self-supervised learning (SSL) models have become one of the most popular methods for the exploration of dysarthric speech recognition due to their much lower demand for labelled data and good ASR results. wav2vec 2.0[12], as a frontrunner SSL model, has demonstrated excellent performance in recognising and transcribing dysarthric speech[13, 14, 15], among which researches [13] and [15] have further validated the effectiveness of including healthy speech in the fine-tuning dataset.

This thesis aims to further explore the effect of combining age-matched healthy speech with dysarthric speech as training dataset during the fine-tuning phase of the a wav2vec 2.0 model to enhance ASR performance for English dysarthric speech. Previous studies have shown the effectiveness of including healthy speech in fine-tuning datasets[13, 15] for dysarthric speech recognition; however, the specific benefit of age-matched healthy speech has not been thoroughly investigated. By experimenting with multiple settings of age-matched healthy speech with dysarthric speech and age-unmatched (general-age) healthy speech with dysarthric speech as a fine-tuning dataset, this research seeks to

[1]Cloud Speech-to-Text V2 documentation, https://cloud.google.com/speech-to-text/v2/docs

[2]The TORGO database - Computer Science, https://www.cs.toronto.edu/ complingweb/data/TORGO/torgo.html

[3]UASpeech, https://ieee-dataport.org/documents/uaspeech

investigate whether this fine-tuning strategy can yield better ASR outcomes.

1.1 Research Question and Hypothesis

In light of the preceding discussion, the research questions at the core of this study can be formulated as follows:

> **1. Does incorporating age-matched healthy speech, instead of general-age healthy speech, improve the fine-tuning performance of the self-supervised model wav2vec 2.0 for English dysarthric speech ASR?**
>
> **2. If the improvement in the first research question is validated, does this improvement also hold in speaker-dependent fine-tuning?**

Since no one has yet attempted to experiment with the inclusion of age-matched healthy speech for fine-tuning self-supervised models specifically for dysarthric speech recognition, my hypothesis for this research question is based primarily on the experimental results from studies [13] and [15], which demonstrated the benefits of including general-age healthy speech. Therefore, my hypothesis for the first research question is that incorporating age-matched healthy speech in the fine-tuning phase can provide more relevant phonetic and acoustic characteristics, leading to better performance of dysarthric speech recognition. Also for the second research question, I hypothesise that this improvement still holds in speaker-dependent fine-tuning.

To validate these hypotheses, I conducted two experiments. Firstly, I fine-tuned the wav2vec 2.0 base checkpoint (pre-trained on English datasets) using three different fine-tuning dataset settings: all dysarthric speech data, a mix of dysarthric and age-matched healthy speech data, and a mix of dysarthric and general-age healthy speech data. The performance of these models was compared using the Word Error Rate (WER) as the evaluation metric. Secondly, I took the best-performing model from the first experiment and re-fine-tune it for speaker-dependent scenarios to further validate the hypothesis.

1.2 Research Contribution

While self-supervised learning models have made significant progress in the field of ASR, no research has yet attempted to fine-tune these models using age-matched healthy speech for dysarthric speech recognition. This research fills this research gap by systematically investigating the effects of incorporating age-matched healthy speech on ASR performance when fine-tuning wav2vec 2.0 models.

By conducting experiments with datasets that include age-matched healthy speech, this research provides theoretical support and practical insights for optimisation strategies of ASR models dealing with dysarthric speech. This approach will not only help to improve the accuracy of speech recognition for patients with dysarthria but also offer reference and guidance for future related research.

The results of this study demonstrates how to maximise the use of limited dysarthric speech data while obtaining great performance, thereby improving the adaptability of the ASR system for dysarthric patients. This research contributes to making speech technology more inclusive and beneficial to user groups with special needs.

1.3 Thesis Outline

The structure of this thesis is as follows: The Introduction section(1) provides the background, motivation, research questions, and hypotheses of the study and outlines the overall structure of the thesis. The Literature Review chapter(2) reviews existing research related to speech recognition for dysarthric speech, covering general knowledge about dysarthria and dysarthric speech, approaches for dysarthric speech recognition with a focus on the self-supervised learning methods, and a thorough explanation of the wav2vec 2.0 model. The Methodology chapter(3) describes the research design and methodology in detail, including model and dataset selection, experimental setting, evaluation metrics, and hardware details. The Results chapter(4) presents the results in the experiment 1 and experiment 2, comparing the ASR performance of models fine-tuned with different datasets. The Discussion chapter (5) interprets the experimental results, validating the first and second experiments, discussing the potential reasons behind the outcomes, pointing out study limitations, and making recommendations for future research. The Conclusion chapter (6) summarises the research objectives and main findings, highlights the contributions and practical applications of the study, and presents a final research outlook.

2 Literature Review

This section provides a comprehensive review of existing research on ASR for dysarthric speech, focusing on self-supervised approaches, with particular attention to the impact of age on dysarthric speech recognition. By conducting a thorough and critical analysis of the literature in this field, this review aims to offer a state-of-the-art overview of the advancements and challenges in dysarthric speech recognition. This comprehensive review not only provides valuable insights into understanding the current state of ASR technology for dysarthria but also highlights the key areas for future research and development.

To those ends, this chapter is structured as follows. First, I provide a brief introduction to dysarthria and dysarthric speech. Next, I explore the popular methods used for dysarthric speech recognition, with a particular focus on self-supervised approaches. Finally, I discuss how age impacts dysarthric speech recognition.

2.1 Dysarthria and Dysarthric Speech

Dysarthria is a motor speech disorder caused by abnormalities in the central or peripheral nervous system, leading to the articulator muscles being paralysed, weakened, or poorly coordinated. These impairments affect several speech productions, including respiration, resonance, articulation, and prosody [16]. Consequently, dysarthria can severely impair functional communication and significantly reduce quality of life [17]. For instance, respiratory issues may lead to reduced breath support, affecting the loudness and length of utterances. Phonatory problems can result in harsh, strained, or breathy voice qualities due to impaired control over vocal fold vibration. Articulatory difficulties often manifest as imprecise consonants and distorted vowels, resulting from reduced tongue, lip, and jaw coordination. Resonance issues, such as hypernasality or hyponasality, occur when the velopharyngeal port does not function properly. Prosodic abnormalities, including monopitch, monoloudness, and inappropriate stress patterns, further complicate the intelligibility of speech[18].

The characteristics of dysarthric speech vary greatly between individuals and are influenced by the type and severity of neurological damage or disease. For instance, damage to the glossopharyngeal nerve affects vocal fold vibration control, leading to a hoarse or rough voice. Vagus nerve damage can impair soft palate movement, causing hypernasality, or excessive nasal airflow during speech[19]. More common problems in speakers with dysarthria include reduced tongue and lip dexterity, resulting in severely slurred speech and difficulty reaching precise articulatory positions for vowel sounds [20]. Additionally, poor articulatory control often produces involuntary sounds due to pharyngeal or vocal fold noise, as well as due to swallowing difficulties [21].

A common feature of different types of dysarthric speech is slowness. In severe cases, dysarthric speech may be up to 17 times slower than normal speech, averaging about 15 words per minute [22]. This slowness makes communication more laborious for both the speaker and listener. For example, listeners often perceive lengthened vowels and prolonged occlusions before voiceless consonants (also known as increased voice onset time or VOT) as breaks or divisions within words or between syllables, which can lead to confusion and misunderstanding of the intended message [23].

Disfluencies are another notable characteristic of dysarthric speech. These include hesitations (e.g., filled pauses) and repetitions (e.g., stuttering), which are especially common when dysarthria is accompanied by aphasia. Although these disfluencies stem from higher-level language problems [24], they result in greatly abnormal phrases that make comprehension at the sentence level difficult.

However, it is important to note that individuals with dysarthria can exhibit considerable variation in their speech patterns. Some individuals with dysarthria, particularly those without accompanying aphasia, may maintain a relatively normal rate of speech and demonstrate more consistency in their speech production. "Relatively normal consistency" refers to their ability to repeat individual speech units (e.g., syllables or words) with more uniform timing and articulation, which can make their speech more understandable compared to those who exhibit frequent disfluencies [20].

Other common speech characteristics of dysarthria include slurred speech, abnormal rhythmic and intonation patterns, and changes in voice quality such as rough or breathy speech sounds. This speech disorder not only affects speakers' ability to communicate effectively but also their social interactions and quality of life.

Furthermore, individuals with dysarthria often experience speaking fatigue, which can exacerbate their speech difficulties over time [25]. The severity of dysarthria can deteriorate progressively, leading to worsening speech production capabilities [25, 26]. There is also high variability between individuals in terms of speech patterns and severity. This variability is influenced by the underlying neurological condition and individual differences in compensatory mechanisms [27].

Altogether, dysarthria not only impairs speakers' ability to communicate effectively but also negatively impacts their social interactions and overall quality of life.

2.2 Dysarthric Speech Recognition

Recognising dysarthric speech presents significant challenges, primarily due to the scarcity of available data and the variability inherent in speech disorders. Traditional supervised learning models, which require large amounts of labelled data, struggle with this limitation. To address this issue, researchers have employed methods such as transfer learning and data augmentation.

Transfer learning, which involves leveraging pre-trained models on large datasets of healthy speech and fine-tuning them with smaller datasets of dysarthric speech, has shown substantial improvements in ASR performance. Techniques such as deep neural network-hidden Markov model (DNN-HMM) hybrid models and end-to-end (E2E) models like Listen, Attend, and Spell (LAS) and Recurrent Neural Network Transducer (RNN-T) have demonstrated that fine-tuning with even a small amount of dysarthric speech data can significantly reduce word error rates (WER) [28, 29, 30, 31, 32, 33].

Data augmentation is another critical technique used to enhance the training datasets by generating synthetic variations of existing data. Methods such as vocal tract length perturbation, tempo perturbation, and speed perturbation have been particularly effective, with studies showing improved ASR performance when these techniques are applied [34, 35, 36].

Despite their effectiveness, these supervised approaches have limitations, such as the need for substantial initial labelled data and the potential introduction of unnatural variations through data augmentation. These challenges highlight the need for alternative methods that can better handle the variability and data scarcity inherent in dysarthric speech recognition.

2.2.1 Self-supervised Learning for Dysarthric Speech Recognition

Self-supervised learning (SSL) has emerged as a promising approach to address the challenges in dysarthric speech recognition, particularly the scarcity of labelled data and the variability in speech patterns. SSL models leverage large amounts of unlabelled data to learn useful speech representations, which can then be fine-tuned with smaller labelled datasets. This section reviews various SSL approaches applied to dysarthric speech recognition.

SSL approaches such as wav2vec 2.0 [12], wav2vec 2.0 XLSR[37], WavLM[38], and HuBERT[39] have shown potential in improving ASR performance for dysarthric speech. These models are pre-trained on large, unlabelled speech corpora using tasks that do not require manual labelling, such as predicting masked parts of the audio signal.

wav2vec 2.0 and Its Variants

wav2vec 2.0 [12] is a popular SSL model that has been extensively studied for dysarthric speech recognition. It consists of a multi-layer convolutional neural network (CNN) feature encoder and a transformer-based context network. During pre-training, the model learns to predict discrete latent speech units from masked audio segments, allowing it to capture rich acoustic representations from unlabelled data [40].

Recent studies have explored the use of wav2vec 2.0 for dysarthric speech. For instance, wav2vec 2.0 and its cross-lingual variant, wav2vec 2.0 XLSR, were evaluated on the UASpeech corpus, showing significant improvements in word error rates (WER) compared to models without SSL pre-training [7]. Experiments demonstrated that incorporating data augmentation techniques, such as GAN-based adversarial perturbation, further enhanced the robustness of these models.

HuBERT

HuBERT (Hidden-Unit BERT) is another SSL model that has been applied to dysarthric speech recognition. HuBERT uses a similar pre-training strategy to wav2vec 2.0 but employs a clustering step to create pseudo-labels, which are then used to train the model to predict masked units. This approach allows HuBERT to learn high-quality speech representations without labelled data [40].

Studies have shown that HuBERT, after fine-tuning with dysarthric speech data, can achieve substantial WER reductions. For example, fine-tuning HuBERT with adversarial data augmentation techniques resulted in improved recognition accuracy for dysarthric speech compared to models trained without such augmentation [40].

WavLM

WavLM (Waveform-based Language Model) is an advanced self-supervised learning (SSL) model designed for a wide range of speech processing tasks, including Automatic Speech Recognition (ASR), speech enhancement, and speaker recognition. Building on the architecture of wav2vec 2.0, WavLM incorporates several enhancements that improve its versatility and performance[38].

WavLM employs a Transformer-based architecture and integrates multiple self-supervised tasks during pre-training, such as masked speech prediction and contrastive learning, to learn robust speech representations. It uses advanced data augmentation techniques, including various noise types and speech perturbations, to improve robustness. Additionally, WavLM incorporates relative positional encodings, enhancing its ability to capture the sequential nature of speech data without the limitations of fixed positional encodings.

One of the key strengths of WavLM is its ability to be pre-trained on large-scale unlabelled speech datasets, allowing it to learn general speech characteristics and patterns. Specifically for WavLM, the integration of diverse noise types and perturbations during pre-training helps the model generalise better to various real-world speech conditions. During fine-tuning, WavLM is adapted to specific tasks or datasets, such as dysarthric speech recognition, using smaller labelled datasets to refine its parameters and achieve high performance in targeted applications[38].

Previous studies using SSL-pretrained ASR models for dysarthric speech recognition have shown promise in terms of improving recognition performance (e.g., WER) [14, 13, 41, 42, 43, 44]. These studies have mostly focused on English dysarthric speech, with varying degrees of success on benchmark datasets like the UASpeech challenge [45]. To investigate the performance of self-supervised pretraining frameworks for dysarthric speech recognition, Violeta et al. [13] examined the wav2vec 2.0 and WavLM models, achieving an average WER of 71.7% and 51.8%, respectively, on the UASpeech test set for 16 dysarthric speakers. Dysarthric speech recognition using the cross-lingual XLSR model was investigated by Hernandez et al. [14], yielding average WERs of 62.0% and 28.6% on the extremely poor and low intelligibility subsets of UASpeech. Furthermore, a study by Shujie Hu [42] examined methods for incorporating SSL pre-trained models and their features into in-domain dysarthric speech trained ASR systems, achieving an average WER of 22.83% on UASpeech, with 52.53% and 25.00% on the extremely poor and low intelligibility subsets.

The results from these studies indicate the potential of SSL models to significantly improve ASR performance for dysarthric speech. However, there remains a significant research gap in the current studies on fine-tuning SSL models for dysarthric speech recognition. Many studies have simply incorporated available healthy speech into their fine-tuning datasets along with dysarthric speech, without exploring the specific characteristics of the healthy speech used. The healthy speech data employed has generally been generic and not tailored to specific needs. However, age-matched healthy speech could potentially provide greater benefits compared to general-age healthy speech [46, 47]. The next subsection reviews the literature on the impact of age on speech recognition and its potential benefits for improving ASR systems.

2.3 Age Helps Dysarthric Speech Recognition

The factor of a speaker's age has a significant influence on the characteristics of dysarthric speech, impacting both the intelligibility and the perceived quality of speech. Understanding how speaker's age interacts with healthy and dysarthric speech patterns is crucial for developing effective ASR systems tailored to different age groups.

Research indicates that ageing affects various acoustic properties of speech, such as fundamental frequency (F0), formant frequencies (F1, F2, F3), and voice onset time (VOT) between younger and older adults [46, 47, 48]. For example, F0 generally increases with age in men but decreases in women, reflecting anatomical and hormonal changes. These changes impact ASR performance as models trained on speech data from younger adults may not generalise well to older adults, who exhibit different pitch patterns and variabilities. Additionally, older adults typically speak more slowly due to reduced cognitive processing speed and neuromuscular changes affecting the precision and speed of articulator movements [46]. This slower speech rate and increased variability in speech production can hinder ASR performance by making it more challenging to accurately segment and transcribe spoken words [47].

Ageing also impacts lexical retrieval, where older adults experience more tip-of-the-tongue (TOT) states and are generally slower and less accurate in naming objects compared to younger adults [46]. This difficulty leads to more filler words and repetitions, which can confuse ASR systems and result in higher word error rates (WER). These age-related changes in speech production have significant implications for ASR performance. Increased variability in acoustic properties, slower speech rates, and lexical retrieval difficulties all contribute to challenges in maintaining high recognition accuracy. Therefore, ASR systems need to be trained on age-matched data to better capture these variations and improve robustness.

In dysarthric speakers, age-related changes in the speech production system can exacerbate existing speech production difficulties [49]. Natural physiological changes, such as reduced lung capacity, decreased muscle strength, and changes in vocal fold elasticity, alter speech rate, pitch, and volume, impacting speech intelligibility [50]. Older individuals with dysarthria may exhibit more pronounced speech rate reductions and increased variability in speech production compared to younger dysarthric speakers[49], making it harder for ASR systems to accurately recognise and transcribe their speech.

The acoustic properties of dysarthric speech, such as lower fundamental frequency and narrower pitch range in older speakers, pose additional challenges for ASR systems not trained on age-matched data[50]. These systems may exhibit higher WER when processing speech with inconsistent loudness, frequent voice breaks, and tremors, which are more common in older dysarthric speakers.

To improve the performance of ASR systems for dysarthric speech across different age groups, it is essential to consider age-related acoustic and perceptual differences. Training ASR models on age-matched data can help capture the specific speech patterns associated with different age groups, leading to more accurate and reliable speech recognition outcomes. Moreover, incorporating advanced acoustic modelling techniques that account for age-related changes in speech production can enhance the adaptability of ASR systems. For example, a dynamic adjustment of model param-

eters based on the estimated age of the speaker could help better handle the variability in speech characteristics observed in older dysarthric individuals.

Despite the promising results from existing research, there remains a significant gap in exploring the use of age-matched healthy speech for fine-tuning ASR models specifically for dysarthric speech recognition. This study aims to address this gap by systematically investigating the impact of incorporating age-matched healthy speech into the fine-tuning datasets of self-supervised models. By doing so, this research seeks to improve ASR performance for older dysarthric speakers, thereby enhancing the inclusivity and effectiveness of speech recognition technology.

3 Methodology

This chapter outlines the methodology employed to address the research questions and validate the hypotheses. The methodology is structured to ensure a comprehensive and systematic approach to model selection, dataset utilisation, experimental setups, evaluation metrics and ethical considerations.

First, in subsection 3.1, the selection of an appropriate model and the explanation of the model are discussed. Next, subsection 3.2 covers the datasets used for training and testing the models, detailing the pre-training, fine-tuning, and evaluation datasets. Subsection 3.3 elaborates on the experimental setups, including data preprocessing, experiment design, and hyperparameter settings. Subsection 3.4 then describes the evaluation method and metrics employed, including Word Error Rate (WER) and Character Error Rate (CER) and also subsequent statistical analysis. Subsection 3.5 provides insights into the hardware and training time considerations, ensuring the reproducibility and feasibility of the experiments. Finally, subsection 3.6 addresses the ethical considerations associated with the research, including data collection, evaluation practices and transparency.

3.1 Model

For this research, the selection of an appropriate model is crucial to ensure efficient use of computational resources and achieve reliable results. The wav2vec 2.0 model, developed by Facebook AI Research, is a powerful self-supervised learning model for speech recognition that has gained significant attention due to its ability to leverage large amounts of unlabelled data for pre-training[12]. This model excels in extracting robust speech representations, which are crucial for tasks like automatic speech recognition (ASR), especially in low-resource settings.

The effectiveness of wav2vec 2.0 is highlighted by its performance on various benchmarks. When pre-trained on large unlabelled datasets like LibriSpeech (960 hours) or LibriVox (60,000 hours), the model significantly reduces word error rates (WER) on standard test sets, even with minimal labelled data for fine-tuning. For example, wav2vec 2.0 achieves a WER of 4.8% on the LibriSpeech test-other set using only 10 minutes of labelled data. Additionally, this model's versatility extends beyond ASR to tasks like speech emotion recognition and speaker identification, showcasing the robustness of its learned representations.

Facebook AI Research has released two main versions of the wav2vec 2.0 model: Base and Large. Each version has multiple pre-trained variants, derived from different datasets, totalling up to 24 pre-trained models. The Base version, which contains 95 million parameters, has shown excellent performance in various tasks such as speech emotion recognition, speaker verification, spoken language understanding, and speech recognition in low-resource languages.

In the current study, I chose to utilise the wav2vec 2.0 Base LS-960 model version for several reasons. First, considering time and computational resource constraints, the Base version is more efficient to work with due to its smaller size compared to the Large version. Additionally, the demonstrated effectiveness of the Base version in handling diverse speech recognition tasks efficiently

made it a suitable choice. The LS-960 variant, pre-trained on the LibriSpeech 960-hour dataset, balances performance and resource efficiency by reducing the number of parameters while maintaining excellent performance[12], making it an ideal candidate for the experiments conducted in this study.

To further elaborate on the capabilities of the wav2vec 2.0 model, it is essential to understand its underlying architecture and the training objectives that contribute to its success. The next subsections delves into the 'Model Architecture' and 'Training Objective', providing a comprehensive overview of how wav2vec 2.0 operates and achieves its remarkable performance.

3.1.1 Model Architecture

The wav2vec 2.0 model is composed of three main parts[12]: a feature encoder, a transformer and a quantisation module.

The feature encoder converts raw audio input into latent speech representations. It consists of multiple convolutional layers that process the input waveform X into a sequence of latent representations $Z = (z_1, z_2, \ldots, z_T)$ for T time steps. Each convolutional layer is followed by a layer normalisation and Gaussian Error Linear Unit (GELU) activation function, ensuring the transformation is non-linear and normalised.

The latent representations Z are then fed into a context network built using the Transformer architecture. The Transformer uses self-attention mechanisms to create contextualised representations $C = (c_1, c_2, \ldots, c_T)$. This mechanism allows the model to capture dependencies across the entire sequence of latent representations, enhancing the overall understanding of the speech context.

For the self-supervised training objective, the output from the feature encoder is discretised into a finite set of speech units via product quantisation. This involves selecting quantised representations from multiple codebooks, concatenating them, and applying a linear transformation. The quantisation is performed using the Gumbel softmax function, which makes the selection process differentiable.

An illustration of the wav2vec 2.0 framework is shown in Figure 1.

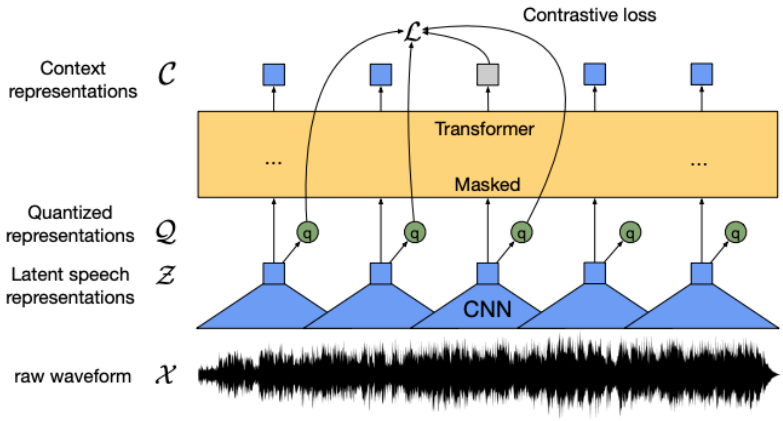

Figure 1: The overview of the wav2vec 2.0 model architecture. Reprinted from [12].

3.1.2 Training Objective

The training of wav2vec 2.0 involves a two-step process[12]: pre-training with self-supervised learning and fine-tuning with labelled data.

During pre-training, the model learns speech representations by masking a random but capped proportion of the latent speech representations and solving a contrastive task. The objective is to identify the true quantised latent representation q_t for a masked time step t among a set of distractors. This task is formalised as a contrastive loss L_m:

$$L_m = -\log \frac{\exp(\text{sim}(c_t, q_t)/\kappa)}{\sum_{\tilde{q} \in Q_t} \exp(\text{sim}(c_t, \tilde{q})/\kappa)}$$

Here, $\text{sim}(a,b)$ denotes the cosine similarity between vectors a and b, and κ is a temperature parameter to scale the logits before applying the softmax function. The distractors are uniformly sampled from other masked time steps in the same utterance.

To ensure the model uses the quantised codebook representations effectively, a diversity loss L_d is included, which maximises the entropy of the averaged softmax distribution over the codebook entries for each codebook group g:

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} p_{g,v} \log p_{g,v}$$

where $p_{g,v}$ represents the probability of selecting the v-th entry in the g-th codebook.

After pre-training, the model is fine-tuned on labelled data using a Connectionist Temporal Classification (CTC) loss to align the predicted sequence with the target transcription. This step optimises

the model for specific speech recognition tasks.

3.2 Datasets

A variety of datasets were employed in this study to ensure a comprehensive evaluation of the ASR models. These datasets include pre-training, fine-tuning, and evaluation datasets, each serving a distinct purpose in the experimental setup. Below, I provide detailed descriptions of each dataset used in this research.

3.2.1 Pre-Training Dataset: LibriSpeech

The primary dataset used for pre-training the wav2vec 2.0 model is the LibriSpeech corpus[51]. LibriSpeech is a large-scale corpus of read English speech derived from audiobooks that are part of the LibriVox project. The dataset contains approximately 1,000 hours of speech sampled at 16 kHz and is publicly available for training and evaluating speech recognition systems.

LibriSpeech was created by aligning the audio books' recordings with their corresponding texts and splitting them into shorter segments suitable for ASR training. The corpus is structured into different subsets based on the quality and difficulty of the recordings. The wav2vec 2.0 Base LS-960 model checkpoint used in this study is pre-trained on the 960-hour subset of the LibriSpeech corpus, providing a diverse and extensive foundation for fine-tuning with dysarthric speech data.

3.2.2 Fine-Tuning, Re-Fine-Tuning and Evaluation Dataset: TORGO

For the fine-tuning and evaluation phases, the TORGO dataset[52] was selected. There are three major datasets available for English dysarthric speech – TORGO, UASpeech[45] and Nemours[5]– with the first two being widely utilised in ASR research. However, due to the unavailability of UASpeech from official channels currently, the TORGO dataset is used exclusively in this study.

The TORGO database is a comprehensive resource designed to support the development of advanced models for dysarthric speech recognition. This dataset includes aligned acoustic and articulatory data from seven individuals with speech impairments caused by cerebral palsy or amyotrophic lateral sclerosis, as well as age- and gender-matched control subjects. Each individual with speech impairments underwent standardised assessments of speech-motor function by a speech-language pathologist.

Acoustic data was collected using one head-mounted and one directional microphone. Articulatory data was obtained through electromagnetic articulography, allowing precise measurement of various aspects of tongues and other articulators' movements during speech, including their trajectories, variability, speed, and correctness of articulation points. Additionally, 3D reconstruction from binocular video sequences was used. The stimuli for the dataset were sourced from various materials, including the TIMIT database, lists of identified phonetic contrasts, and assessments of speech

intelligibility.

This dataset also includes analyses of how dysarthric speech differs from non-dysarthric speech based on features such as phoneme length and pronunciation errors, as described in its official documentation[52]. By providing both acoustic and detailed physiological information, the TORGO dataset enables the development of ASR systems that can better handle the unique challenges posed by dysarthric speech.

For evaluation purposes, specific samples were carefully selected from the TORGO dataset to ensure that no content overlapped between the fine-tuning and evaluation datasets. This careful selection process helps to maintain the integrity of the evaluation by preventing any potential bias that could arise from the models being exposed to the same content during both fine-tuning and evaluation phases.

3.2.3 Control Fine-tuning Dataset: Common Voice Delta Segment 17.0

The Common Voice project[4] is designed to incorporate as many languages as possible for the development of inclusive speech recognition systems[53]. Recordings are made by community contributors who read given prompts. These recordings are then validated by other contributors, which helps to alleviate the need for expert transcription. The number of hours of speech data available can vary depending on the version of the corpus.

For this study, the Common Voice Delta Segment 17.0 dataset was used. This version contains 1.6 GB of data, with 70 recorded hours and 30 validated hours. The choice of the Common Voice dataset as the source of general-age healthy speech is intentional and strategic. Unlike other healthy English speech datasets, Common Voice offers a broad and diverse demographic representation from 1,851 voices, which is essential for testing the robustness and generalisability of the ASR models. By comparing this dataset with age-matched healthy speech from the TORGO dataset, I assess the effectiveness of incorporating age-matched healthy speech into the fine-tuning process. This comparison allows me to determine whether the inclusion of age-matched healthy speech provides a significant advantage over a more varied, demographically diverse dataset like Common Voice in improving the ASR system's performance for dysarthric speech.

3.3 Experimental Settings

3.3.1 Data Preprocessing

The preprocessing of the TORGO and Common Voice datasets involved several critical steps to ensure data quality and consistency.

For the TORGO dataset, which includes recordings made using two different microphones (an array microphone and a head-worn microphone) capturing the same utterances, the array microphone

[4]Common Voice, https://commonvoice.mozilla.org/cv/datasets

recordings were chosen due to their higher clarity and lower electrical noise after sampling and listening to various recordings[52]. Any recordings with significant discrepancies between the speech and transcription, such as background speech or repeated utterances, were removed.

Transcription normalisation was performed by converting all transcriptions to lowercase letters and removing all punctuation marks. This step was particularly important for standardising the data, as dysarthric speech often includes irregular pauses and stuttering, which can lead to varied punctuation usage. By removing punctuation and converting to lowercase, the text data becomes more uniform, allowing the ASR model to focus on the core phonetic content without being misled by inconsistencies in transcription.

In the first experiment, the TORGO dataset was then split into training, validation, and evaluation sets. The total dysarthric speech amounted to approximately 200 minutes, which was divided into 140 minutes for training, 30 minutes for validation, and 30 minutes for evaluation. Additionally, there were 250 minutes of age-matched healthy speech from TORGO and 250 minutes of general-age healthy speech from Common Voice. It was ensured that the transcriptions in the validation and evaluation datasets did not appear in the training dataset, across all types of speech (dysarthric, age-matched healthy, and general-age healthy).

In the second experiment, high-intelligibility speech data from the speaker labelled F03 and low-intelligibility speech data from the speaker labelled M04, along with their corresponding age-matched healthy speech, were extracted from the fine-tuning data used in the first experiment, resulting in 40 minutes from F03 with 40 minutes from FC03 (female control group), and 40 minutes from M04 with 40 minutes from MC04 (male control group) respectively. This allowed for a focused analysis of the impact of age-matched healthy speech on speaker-dependent fine-tuning.

For the Common Voice dataset, recordings were meticulously reviewed to exclude any samples with significant background noise or poor transcription alignment. Similar to the TORGO dataset, all transcriptions were standardised by converting them to lowercase and removing punctuation to maintain consistency in preprocessing. In all experimental conditions, the duration of general-age healthy speech was matched to the duration of age-matched healthy speech to ensure uniformity in the amount of training data used for fine-tuning the models.

3.3.2 Experiment Design

This research designs two experiments. The first experiment was speaker-independent. Speech data from all dysarthric speakers were mixed together, and similarly, speech data from age-matched healthy speakers and general-age healthy speakers were also mixed. The wav2vec 2.0 Base checkpoint was then fine-tuned using three different datasets: only dysarthric speech, age-matched healthy speech combined with dysarthric speech, and general-age healthy speech combined with dysarthric speech. The performance of the three fine-tuned models was then compared using the same evaluation set.

In the second experiment, which followed the first, the best-performing checkpoint from the first experiment was used for further fine-tuning, but this time the experiment was speaker-dependent.

Speech from two dysarthric speakers with varying severity levels (low and high intelligibility) was used. Both of these speakers' data was used to re-fine-tune the model separately with age-matched healthy speech and general-age healthy speech. This resulted in four models for comparison, and their performance was later compared (see subsection 4.2). Also, in the first experiment, the model fine-tuned with only dysarthric speech performed the worst, as the other two sets included additional healthy speech data. Therefore, in the second experiment, the only dysarthric speech setting was not used for fine-tuning.

3.3.3 Hyperparameters Setting

In the first experiment, multiple trials were conducted to determine the optimal hyperparameters for stable model convergence and performance. The following settings were found to be the most effective:

- Batch size: 8
- Optimiser: torch.optim.AdamW
- Learning rate (lr): 1e-5
- Scheduler: torch.optim.lr_scheduler.OneCycleLR
- Maximum learning rate (max_lr): 1e-4
- Validation interval: 500
- Gradient accumulation steps: 2

For the second experiment, these parameters were kept consistent. However, due to the relatively small dataset size, K-fold cross-validation was introduced to ensure robustness. The K-fold value was set to 5, allowing the model to train and validate on different data splits, which allowed for the generalisation and reliability of the results.

3.4 Evaluation and Analysis

3.4.1 Evaluation Metrics

The evaluation of the models in this study was primarily based on the Word Error Rate (WER), a widely used metric in the field of automatic speech recognition. WER is defined as:

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

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference text. A lower WER indicates better performance of the model. WER is expressed as a percentage and can exceed 100%, though it has a lower boundary of 0%.

Additionally, the Character Error Rate (CER) was used as an auxiliary metric to provide a more granular view of the model's performance, especially in cases where individual character accuracy is critical. CER is calculated similarly to WER but operates at the character level:

$$\text{CER} = \frac{S + D + I}{N}$$

where S is the number of incorrect characters, D is the number of missing characters, I is the number of extra characters, and N is the total number of characters in the reference text.

For the evaluation, the results from the test set are analysed using both WER and CER. This dual-metric approach provides a comprehensive assessment of the model's accuracy in recognising dysarthric speech. The comparison of models from each experiment involves looking at absolute and relative differences in WER and CER, allowing for a detailed performance analysis across different fine-tuning strategies.

3.4.2 Statistical Analysis

To ensure the robustness and reliability of the evaluation, statistical analysis was performed on the values of the evaluation WER and CER for all test samples in both Experiment 1 and Experiment 2. This analysis was carried out to understand the distribution of the error rates and to validate the significance of the observed differences between the models.

The Shapiro-Wilk test is a powerful statistical test for assessing the normality of data and was employed to test the evaluation WER and CER for normal distribution. It provides a statistic that measures how well the data conforms to a normal distribution, with a corresponding p-value indicating the significance of the result. The Shapiro-Wilk test was used to determine whether parametric or non-parametric statistical tests should be applied in further analysis.

Since the data was not normally distributed (see Results4), the non-parametric alternative of ANOVA, the Kruskal-Wallis test, which does not assume a normal distribution, was used for the group analysis of WER and CER values. It compares the medians of multiple groups to assess whether there are statistically significant differences. The use of the Kruskal-Wallis test ensures that the analysis is robust and valid even when the normality assumption is violated.

By conducting these statistical tests, the evaluation ensures that the differences in performance between models are not only observed but also statistically validated. This rigorous approach enhances the credibility of the findings and supports the conclusions drawn from the experiments.

3.5 Hardware and Training Time

The experiments were conducted on the Hábrók high-performance computing cluster of the University of Groningen. The GPU used was an Nvidia A100 GPU accelerator card with 40 GB of VRAM available.

The following lists show the time taken for fine-tuning in each experimental condition:

Experiment 1:

- Fine-tuning on all dysarthric speech: 2.3 hours
- Fine-tuning on dysarthric and age-matched healthy speech: 4.5 hours
- Fine-tuning on dysarthric and general-age healthy speech: 4.5 hours

Experiment 2:

- Re-fine-tuning for high-intelligibility on dysarthric and age-matched healthy speech: 1.7 hours
- Re-fine-tuning for high-intelligibility on dysarthric and general-age healthy speech: 1.6 hours
- Re-fine-tuning for low-intelligibility on dysarthric and age-matched healthy speech: 1.6 hours
- Re-fine-tuning for low-intelligibility on dysarthric and general-age healthy speech: 1.5 hours

These durations indicate the computational resources required for each experimental condition, providing insight into the feasibility and efficiency of the fine-tuning processes. The relatively short fine-tuning times suggest that the methodology was efficient and can be executed within reasonable time frames, even when re-fine-tuning for specific speaker groups. This efficiency is crucial for practical applications where rapid model adaptation is needed, such as in personalised ASR systems for dysarthric speakers.

3.6 Ethical Considerations

While the primary aim of this research was to develop an ASR system that enhances the accessibility and usability of speech recognition technology for dysarthric speakers, it was crucial to acknowledge and mitigate potential ethical concerns and risks associated with such technology. In this regard, I am committed to transparent communication of the study's results and implications.

3.6.1 Data Collection and Use

The data utilised in this research was derived from previously recorded datasets, including the Mozilla Common Voice project and other publicly available corpora. These datasets are open,

crowdsourced, and continually updated, ensuring a wide range of linguistic diversity. The participants in these projects were informed about their contributions, and their data was collected voluntarily[53, 52]. The Mozilla Common Voice dataset, in particular, is licensed under CC0, allowing free distribution and adaptation without requiring credit[5].

The TORGO database, specifically designed for dysarthric speech research, contains recordings from speakers with various types and severities of dysarthria. The use of this dataset is particularly significant for the research focus on improving ASR for dysarthric speakers. The TORGO database is also publicly available, and participants have consented to their data being used for research purposes[52].

3.6.2 Evaluation Metrics

The evaluation of the ASR models is based on objective metrics such as Word Error Rate (WER) and Character Error Rate (CER), which are standard in the field of speech recognition. Subjective evaluation methods involving human participants have not been employed, thus avoiding potential ethical concerns related to human subject research.

3.6.3 Transparency and Replicability

In line with the principles of open science, all code and models developed during this research are made publicly available via GitHub[6]. Detailed instructions on reproducing the experiments are provided, ensuring that other researchers can validate and build upon this work. While minor variations in outcomes may occur due to hardware differences or inherent randomness in model training, the overall methodology and findings remain consistent.

In summary, this chapter detailed the comprehensive methodology used to develop and evaluate an ASR system for dysarthric speech, encompassing model selection, dataset utilisation, experimental setups, and evaluation metrics. The careful design and rigorous implementation ensure the reliability and reproducibility of the experiments. The following chapter presents the results obtained from these experiments, indicating the performance of the ASR models fine-tuned with different datasets.

[5] common-voice/LICENSE, missuniverse and mikehenrty, https://github.com/common-voice/common-voice/blob/main/LICENSE

[6] https://github.com/CantaoSu/DysarthricASR-FineTuning-by-Different-Datasets

4 Results

This chapter presents the findings from the experiments conducted to investigate the effect of incorporating age-matched healthy speech alongside dysarthric speech during the fine-tuning phase of the wav2vec 2.0 model for English dysarthric speech ASR. The primary focus of this chapter is on comparing the performance of models fine-tuned with different datasets: only dysarthric speech, dysarthric speech with age-matched healthy speech, and dysarthric speech with general-age healthy speech.

The primary evaluation metric is Word Error Rate (WER), which was used to assess the models' accuracy. Since no language model was incorporated, Character Error Rate (CER) was also used as an auxiliary metric to provide additional insights into the models' performance.

4.1 Experiment 1

Experiment 1 aimed to evaluate the performance of the wav2vec 2.0 model when fine-tuned with dysarthric speech alone and in combination with either age-matched healthy speech or general-age healthy speech.

The performance of the models in Experiment 1 is summarised in Table 1:

Table 1: Performance of Models in Experiment 1

Model	Fine-tuning Dataset	Val WER	Evaluation WER	Evaluation CER
Model 1	Only dysarthric speech	80.28	74.96	37.71
Model 2	+ Age-matched healthy	62.24	**47.63**	**23.34**
Model 3	+ general-age healthy	71.96	60.86	30.20

Subsequently, a Shapiro-Wilk test was conducted to test the evaluation WER and evaluation CER of all the evaluation samples for normal distribution. The Shapiro-Wilk test results are summarised in Table 2:

Table 2: Shapiro-Wilk Test Results for Experiment 1

WER	Statistic	p-value
WER_model1	0.63	6.85×10^{-25}
WER_model2	0.72	5.66×10^{-22}
WER_model3	0.68	2.75×10^{-23}

CER	Statistic	p-value
CER_model1	0.94	8.96×10^{-10}
CER_model2	0.89	1.09×10^{-13}
CER_model3	0.91	2.93×10^{-12}

Since all distributions were not normally distributed ($p < 0.05$), a non-parametric Kruskal-Wallis test was used. The Kruskal-Wallis test results are reported in Table 3.

Table 3: Kruskal-Wallis Test Results for Experiment 1

Metric	**H-statistic**	**p-value**
WER	12.45	0.002
CER	12.75	0.002

The Kruskal-Wallis test results show a significant difference in WER and CER among the three models. This indicates that the differences in performance are statistically significant.

To identify which groups are significantly different from each other, Dunn's post hoc test was performed. The results of Dunn's post hoc test for WER and CER are shown in Tables 4 and 5, respectively.

Table 4: Dunn's Post Hoc Test for WER

	Model 1	**Model 2**	**Model 3**
Model 1	1.000	0.001	0.337
Model 2	0.001	1.000	0.159
Model 3	0.337	0.159	1.000

Table 5: Dunn's Post Hoc Test for CER

	Model 1	**Model 2**	**Model 3**
Model 1	1.000	0.001	0.226
Model 2	0.001	1.000	0.219
Model 3	0.226	0.219	1.000

Using a significance level (α) of 0.05[7], the Dunn's post hoc test results show that:

- For WER, the difference between Model 1 and Model 2 is statistically significant ($p = 0.001$), indicating that Model 2 significantly outperforms Model 1. The differences between Model 1 and Model 3 ($p = 0.337$) and between Model 2 and Model 3 ($p = 0.159$) are not statistically significant.
- For CER, the difference between Model 1 and Model 2 is statistically significant ($p = 0.001$), indicating that Model 2 significantly outperforms Model 1. The differences between Model 1 and Model 3 ($p = 0.226$) and between Model 2 and Model 3 ($p = 0.219$) are not statistically significant.

Results from Table 1 confirm that Model 2, fine-tuned with dysarthric speech and age-matched healthy speech, shows the best performance with significantly lower WER and CER compared to

[7]Mastering Kruskal-Wallis and Dunn's Test: A Comprehensive Guide,
https://www.adventuresinmachinelearning.com/mastering-kruskal-wallis-and-dunns-test-a-comprehensive-guide/

Model 1, which was fine-tuned with only dysarthric speech. Model 3, which included general-age healthy speech, showed an improvement over Model 1 but did not match the performance of Model 2.

The results from Experiment 1 indicate that the inclusion of age-matched healthy speech in the fine-tuning dataset improves the model's ability to recognise dysarthric speech. The lower WER and CER for Model 2 suggest that the demographic matching in the training data contributes positively to the model's performance.

4.2 Experiment 2

Experiment 2 focused on speaker-dependent fine-tuning using the best-performing model from Experiment 1 (Model 2) and further fine-tuning it with speech data from dysarthric speakers of varying intelligibility levels, combined with either age-matched healthy speech or general-age healthy speech.

The performance of the models in Experiment 2 is summarised in the Table 6:

Model	Intelligibility	Fine-tuning Dataset	Val WER	evaluation WER	evaluation CER
Model 4	High	+ general-age healthy	65.71	43.27	21.02
Model 5	High	+ Age-matched healthy	8.75	**35.58**	**15.91**
Model 6	Low	+ general-age healthy	81.66	**75.70**	**38.84**
Model 7	Low	+ Age-matched healthy	77.64	77.57	43.82

Table 6: Performance metrics for models in Experiment 2

The results indicate that incorporating age-matched healthy speech improves ASR performance for high-intelligibility dysarthric speakers but shows mixed results for low-intelligibility speakers. For high-intelligibility speakers, Model 5 (incorporating age-matched healthy speech) achieves the lowest evaluation WER (35.58) and CER (15.91), outperforming Model 4 (incorporating general-age healthy speech), which has a evaluation WER of 43.27 and a CER of 21.02.

However, for low-intelligibility speakers, the performance does not follow the same pattern. Model 6 (incorporating general-age healthy speech) performs better with a evaluation WER of 75.70 and a CER of 38.84 compared to Model 7 (incorporating age-matched healthy speech), which has a higher evaluation WER of 77.57 and a CER of 43.82.

A Shapiro-Wilk test was conducted to test the evaluation WER and evaluation CER of all the evaluation samples for normal distribution. The results as the Table 7 indicate that the data were not normally distributed.

Table 7: Shapiro-Wilk Test Results for Experiment 2

WER	Statistic	p-value
WER_model4	0.67	9.64×10^{-9}
WER_model5	0.69	1.86×10^{-8}
WER_model6	0.68	1.02×10^{-8}
WER_model7	0.60	6.31×10^{-10}

CER	Statistic	p-value
CER_model4	0.90	9.03×10^{-4}
CER_model5	0.84	1.97×10^{-5}
CER_model6	0.96	8.44×10^{-3}
CER_model7	0.93	6.14×10^{-3}

Given the non-normality of the data, the non-parametric Kruskal-Wallis test was used to compare the models. The results are as the Table 8 shows:

Table 8: Kruskal-Wallis Test Results for Experiment 2

Intelligibility	H-statistic	p-value
WER_High	1.18	0.28
WER_Low	0.92	0.34

Intelligibility	H-statistic	p-value
CER_High	1.77	0.18
CER_Low	3.45	0.06

The Kruskal-Wallis test results for WER and CER in Experiment 2 indicate that the differences between the models for both high and low intelligibility groups were not statistically significant.

The results from the Table 6 show that Model 5, which incorporated age-matched healthy speech, performs better than Model 4 for high-intelligibility speakers. This suggests that demographic specificity in the training data can enhance the model's accuracy in transcribing high-intelligibility dysarthric speech. For low-intelligibility speakers, Model 6, which used general-age healthy speech, shows slightly better performance than Model 7, which used age-matched healthy speech. This outcome highlights that the benefits of demographic matching are less clear for low-intelligibility dysarthric speech. Further interpretation and implications of these findings are discussed in detail in the Discussion chapter 5.

For a deeper understanding of the models' learning dynamics and to visually track the optimisation process, visualisations of loss and WER changes throughout the fine-tuning phase can be found in Appendix A.

5 Discussion

This chapter provides an interpretation of the experimental results and an analysis of the performance of the wav2vec 2.0 model when fine-tuned with different datasets for English dysarthric speech recognition. It contextualises the findings within the research hypotheses, explores the potential reasons behind the observed outcomes, and acknowledges the limitations of the study. This chapter is essential for understanding the implications of the results and for guiding future research in the field of ASR for dysarthric speech.

The primary focus of this research was to evaluate whether incorporating age-matched healthy speech into the fine-tuning phase of the self-supervised model wav2vec 2.0 would improve its performance in recognising dysarthric speech. Specifically, the research aimed to answer two key questions:

1. Does incorporating age-matched healthy speech, instead of general-age healthy speech, improve the fine-tuning performance of the self-supervised model wav2vec 2.0 for English dysarthric speech ASR?
2. If the improvement in the first research question is validated, does this improvement also hold in speaker-dependent fine-tuning?

The hypotheses corresponding to these questions were that age-matched healthy speech would provide more relevant phonetic and acoustic characteristics, leading to better performance in dysarthric speech recognition, and that this improvement would also hold in speaker-dependent fine-tuning scenarios.

5.1 Validation of the First Hypothesis

The first experiment was designed to test the performance of the wav2vec 2.0 model fine-tuned with different datasets: only dysarthric speech (Model 1), dysarthric speech combined with age-matched healthy speech (Model 2), and dysarthric speech combined with general-age healthy speech (Model 3). The results indicate that Model 2, which includes age-matched healthy speech combined with dysarthric speech for fine-tuning, achieves the best performance.

Both Model 2 and Model 3 outperformed Model 1 which relies solely on dysarthric speech. This improvement can primarily be explained by the increased amount of speech data, which provides more references for the wav2vec 2.0 model to learn high-level speech characteristics.

Despite not being target speech, the additional healthy speech data aids in the model's learning process. Both Model 2 and Model 3 incorporate an equal amount (250 minutes) of healthy speech, yet Model 2 performs better. This suggests that age-matched healthy speech may be more beneficial than general-age healthy speech for enhancing dysarthric speech recognition.

There are at least several possible explanations for such effect. One possible explanation for this improvement is the similarity in speech patterns. Age-matched healthy speech is likely to exhibit

similar speech patterns and characteristics to those found in dysarthric speech, which helps the model in learning more relevant features. Young and Mihailidis (2010)[49] discussed how age-related changes in speech can cause older adults' voices to resemble aspects of dysarthric speech, which supports this explanation.

Another potential reason is the consistency in acoustic features. Age-matched healthy speech may share more consistent acoustic features with dysarthric speech, such as pitch and tone, which enhances the model's ability to generalise. This is supported by Warmbier et al. (2023), who emphasised the importance of age and text type in assessing speech disorders, highlighting how acoustic similarities can impact performance time and accuracy in speech assessments[50].

Furthermore, age-matched speech might provide a contextually richer dataset that allows the model to learn age-specific nuances, leading to better performance. This aligns with the general findings that the age of speakers significantly influences the characteristics of their speech[46, 47, 48], affecting the model's ability to accurately recognise and process spoken words.

Altogether, the findings of the current study show that incorporating age-matched healthy speech data not only enhances the model's ability to learn relevant features but also provides a more consistent and contextually appropriate dataset. This suggests that when selecting data for fine-tuning, one should consider not only the quantity but also the quality and relevance of the additional speech data.

Given these findings, it was expected that the benefits of using age-matched healthy speech would also extend to speaker-dependent fine-tuning, providing more personalised and accurate recognition for individual speakers.

5.2 Validation of the Second Hypothesis

The second experiment, which involved re-fine-tuning for speaker-dependence, yielded varying results. Age-matched healthy speech (Model 5) demonstrated superior performance for high-intelligibility speakers compared to general-age healthy speech (Model 4). This finding aligns with the expectation derived from the first experiment, as the closer age approximation likely offers more similar acoustic characteristics, aiding the model in better generalising the speech patterns.

However, for low-intelligibility speakers, the performance did not support the original hypothesis. Model 7, which used age-matched healthy speech, did not outperform Model 6, which used general-age healthy speech. This discrepancy might be due to the greater inherent variability in the speech patterns of low-intelligibility speakers. Severe dysarthric speech has more unpredictable and complex variations than healthy speech[54, 55, 56, 57], which might not be sufficiently captured by the additional age-matched healthy speech data. Consequently, the benefits of age-matching might be less pronounced for low-intelligibility speech, where diverse and less predictable speech patterns dominate the model's learning process.

These results highlight the importance of considering the specific characteristics of the target speech when selecting fine-tuning data. While age-matched healthy speech can provide substantial benefits,

especially for high-intelligibility speakers, its effectiveness diminishes with the increasing severity of speech impairment, suggesting the need for more tailored approaches in such cases. These findings are supported by the broader body of research in dysarthric speech recognition. For instance, previous studies have demonstrated the effectiveness of various fine-tuning strategies for improving ASR performance on dysarthric speech. Shor et al. (2019)[31] and Green et al. (2021)[30] showed that even small amounts of dysarthric speech data could significantly reduce WER when used to fine-tune models pre-trained on large datasets of healthy speech.

However, these studies often utilised generic healthy speech data without considering age-matched characteristics. By focusing on the inclusion of age-matched healthy speech, my research addresses a gap in the current literature. The evidence suggests that age-matched healthy speech provides more relevant phonetic and acoustic characteristics, leading to better performance in recognising dysarthric speech, especially for high-intelligibility speakers. This aligns with findings from studies on age-related changes in speech[46, 47, 48], which indicate that age-specific data can improve model generalisation and accuracy.

Furthermore, the second experiment underlines the significance of personalised and adaptive approaches in ASR for dysarthric speech. The varying performance across different levels of intelligibility indicates that a one-size-fits-all model may not be the most effective strategy. Instead, adaptive models that can be fine-tuned with highly specific and relevant datasets may offer better performance and usability for individuals with severe speech impairments.

This adaptive approach is supported by the work of Shujie Hu et al.(2023), who demonstrated that incorporating SSL pre-trained models into in-domain dysarthric speech-trained ASR systems could achieve lower WERs by tailoring the fine-tuning process to specific speech characteristics[33, 42]. While my study does not directly compare results with every previous study due to differences in datasets and methodologies, it builds on the established understanding that fine-tuning with relevant data is crucial for improving ASR performance.

5.3 Limitations

There are a few limitations to the study that need to be addressed.

The first limitation concerns the variability in transcription content. The transcription content between age-matched and age-unmatched healthy speech differs significantly. This variability in transcription content, along with the age differences, may have influenced the model's performance during training. Age-unmatched healthy speech contains a somewhat different range of vocabulary and sentence structures compared to age-matched speech, potentially leading to inconsistencies in model training. The Cosine Similarity between the two datasets was calculated to be 0.89, indicating that while the overall vocabulary and structure of the two texts are quite similar, there are still notable differences. The Chi-square Test further supports this, with a chi-square statistic of 14897.97 and a p-value smaller than 0.001, demonstrating significant differences in word frequency distributions between the two texts. Additionally, the Average Sentence Length in the age-unmatched dataset (Common Voice) was 5.84 words per sentence compared to 2.62 words per sentence in the

age-matched dataset, highlighting differences in sentence complexity.

The impact of these differences on model performance could be better understood if age-matched and age-unmatched healthy speech samples with identical content were available, allowing for a more controlled comparison. Given that this study worked with read speech, having such controlled samples would help isolate the effect of age-matching from the effects of vocabulary and sentence structure differences, providing a clearer picture of how these factors influence ASR performance for dysarthric speech.

The second limitation arises from the size and composition of the dataset. The TORGO dataset, used for fine-tuning, is relatively small and includes only a limited number of speakers, particularly those with moderate intelligibility. The dataset comprises seven individuals with dysarthria caused by cerebral palsy or amyotrophic lateral sclerosis and the age- and gender-matched controls. However, the absence of age-matched healthy speech for individuals with moderate intelligibility and the lack of age information for all speakers limit the robustness and generalisability of the findings. Future research should aim to include a larger and more diverse dataset that represents a broader range of dysarthric speech characteristics.

The third limitation is the limited speaker diversity. Due to time constraints, the second experiment only re-fine-tuned and compared models using data from individual speakers with varying levels of intelligibility. This approach did not allow for validation or comparison across other speakers with the same intelligibility level. As a result, the findings may not fully capture the variability present in the broader population of dysarthric speakers. Including a wider range of speakers with similar intelligibility levels in future studies would provide a more comprehensive evaluation of the models' effectiveness.

The fourth limitation is that the fine-tuning dataset, TORGO, only consists of read speech. This does not fully align with the real-world applications of dysarthric speech recognition, where spontaneous speech is more common, such as in smart home environments. Recognising dysarthric speech in everyday settings often involves spontaneous, rather than read, speech. Therefore, the model's performance might differ when applied to spontaneous speech scenarios, highlighting the need for datasets that better reflect the practical usage of dysarthric speech recognition systems.

By addressing these limitations, future research can enhance the robustness and applicability of ASR systems for dysarthric speech. This will ultimately improve their effectiveness in real-world scenarios, ensuring that speech recognition technology is more inclusive and accessible for individuals with speech impairments.

5.4 Future Work

To build on the findings of this study and address its limitations, several avenues for future research are proposed. These suggestions aim to enhance the understanding and performance of ASR systems for dysarthric speech, ensuring broader applicability and clinical relevance.

- Intelligibility Level Verification: A deeper investigation into the relationship between the

severity and intelligibility levels of dysarthric speech and the requirements for age-matched healthy speech is necessary. This may involve fine-tuning models with datasets that represent a broader spectrum of intelligibility, providing a more nuanced understanding of the impact on ASR performance.

- Dataset Expansion: Future studies should consider a wider range of datasets, including those from underrepresented languages and those with more extensive and diverse speaker demographics. This would allow for a more comprehensive assessment of model performance across various speech patterns and linguistic backgrounds.
- Advanced Models Exploration: The application of other cutting-edge models such as WavLM, data2vec 2.0, and other innovative ASR architectures should be pursued. These models may offer novel strategies for improving ASR performance and could potentially yield higher accuracy rates for dysarthric speech recognition.
- Cross-Linguistic Generalisation: Extending the research to other languages would allow to determine the generalisability of the findings and contribute to the development of ASR systems that are effective across linguistic boundaries.
- Data Augmentation and Model Customisation: Further research into advanced data augmentation techniques and the customisation of model architectures is essential. These approaches could better accommodate the diverse characteristics of dysarthric speech and lead to improved ASR performance.
- Clinical Collaboration: Engaging with speech therapists and clinicians to incorporate insights from dysarthria research is vital. This collaboration could guide the development of ASR systems that are not only technologically advanced but also clinically relevant and beneficial for individuals with dysarthria.

6 Conclusions

This research has endeavoured to bridge the gap in ASR technology's ability to effectively serve individuals with dysarthria, focusing on the strategic incorporation of age-matched healthy speech data during the fine-tuning of ASR models.

6.1 Summary of Findings

This study addresses a significant gap in the accessibility and effectiveness of automatic speech recognition (ASR) technology for individuals with dysarthria. The primary objective was to investigate the impact of incorporating age-matched healthy speech data during the fine-tuning phase of the wav2vec 2.0 model for English dysarthric speech ASR. The research provided substantial evidence supporting the benefits of this approach, particularly for high-intelligibility dysarthric speakers.

Experiment 1 aimed to evaluate the performance of the wav2vec 2.0 model when fine-tuned with dysarthric speech alone and in combination with either age-matched healthy speech or general-age healthy speech. The results from Experiment 1 were particularly telling, as models fine-tuned with dysarthric speech combined with age-matched healthy speech consistently outperformed those that were not.

Building upon the initial findings, Experiment 2 delved deeper into the nuances of speaker-dependent fine-tuning. This analysis revealed that while age-matched healthy speech significantly benefits high-intelligibility dysarthric speakers, its effectiveness for low-intelligibility speakers was less clear-cut. These mixed results suggest that the relationship between the age-matched healthy speech and the target dysarthric speech is complex and may be influenced by the severity of the speech impairment.

6.2 Main Contributions

A critical contribution of this research lies in its methodological framework. The research established a systematic approach to fine-tuning ASR models using dysarthric speech datasets. This methodology encompasses model selection, dataset preparation, experimental design, and evaluation metrics, providing a replicable and comprehensive framework for future research in this domain.

Moreover, this research contributes to the broader goal of making speech technology more inclusive. By demonstrating the potential benefits of demographically matched data in ASR systems, the research has paved the way for more personalised and effective ASR solutions for users with special speech needs. The findings underscore the importance of considering the unique characteristics and requirements of different user groups when developing and fine-tuning ASR models. This research has not only advanced the scientific understanding of ASR for dysarthric speech but has also taken steps towards ensuring that technological advancements in this field are inclusive and equitable.

Bibliography

[1] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[2] Lyle Campbell and Raymond G. Gordon. *Language*, 84(3):636–641, 2008.

[3] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey, 2021.

[4] Ray Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. *Ending Parkinson's disease: a prescription for action*. Hachette UK, 2020.

[5] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[6] Saeid Alavi Naeini, Leif Simmatis, Deniz Jafari, Yana Yunusova, and Babak Taati. Improving dysarthric speech segmentation with emulated and synthetic augmentation. *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.

[7] Ryoichi Takashima, Yuya Sawa, Ryo Aihara, Tetsuya Takiguchi, and Yoshie Imai. Dysarthric speech recognition using pseudo-labeling, self-supervised feature learning, and a joint multi-task learning approach. *IEEE Access*, 2024.

[8] Emre Yılmaz, Vikramjit Mitra, Ganesh Sivaraman, and Horacio Franco. Articulatory and bottleneck features for speaker-independent asr of dysarthric speech. *Computer Speech amp; Language*, 58:319–334, November 2019.

[9] Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini, and Helmer Strik. Multi-Stage DNN Training for Automatic Recognition of Dysarthric Speech. In *Proc. Interspeech 2017*, pages 2685–2689, 2017.

[10] Emre Yılmaz, Mario Ganzeboom, Catia Cucchiarini, and Helmer Strik. Combining Non-Pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech. In *Proc. Interspeech 2016*, pages 218–222, 2016.

[11] Emre Yılmaz, Vikramjit Mitra, Chris Bartels, and Horacio Franco. Articulatory features for asr of pathological speech, 2018.

[12] 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.

[13] Lester Phillip Violeta, Wen-Chin Huang, and Tomoki Toda. Investigating self-supervised pre-training frameworks for pathological speech recognition. *arXiv preprint arXiv:2203.15431*, 2022.

[14] Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition, 2022.

[15] Tatsunari Matsushima. *Dutch dysarthric speech recognition: Applying self-supervised learning to overcome the data scarcity issue*. PhD thesis, 2022.

[16] Pam Enderby. Disorders of communication: dysarthria. *Handbook of clinical neurology*, 110:273–281, 2013.

[17] Lena Hartelius, Marie Elmberg, Rebecca Holm, Ann-Sofie Lövberg, and Stiliani Nikolaidis. Living with dysarthria: evaluation of a self-report questionnaire. *Folia phoniatrica et logopaedica*, 60(1):11–19, 2008.

[18] Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269, 1969.

[19] Bianca Maria Liquidato and Feres Chaddad Neto. Glossopharyngeal schwannoma causing vocal fold paralysis. *Revista Brasileira de Otorrinolaringologia*, 74:947–947, 2008.

[20] Ray D Kent and Kristin Rosen. Motor control perspectives on motor speech disorders. *Speech motor control in normal and disordered speech*, pages 285–311, 2004.

[21] Kristin Rosen and Sasha Yampolsky. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16(1):48–60, 2000.

[22] Rupal Patel, Christopher Dromey, and Hans Kunov. Control of prosodic parameters by an individual with severe dysarthria. *Univ. of Toronto, Toronto, ON, Canada, Tech. Rep*, 1998.

[23] Parimala Raghavendra, Elisabet Rosengren, and Sheri Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication*, 17(4):265–275, 2001.

[24] Ray D Kent. Research on speech motor control and its disorders: A review and prospective. *Journal of Communication disorders*, 33(5):391–428, 2000.

[25] Shansong Liu, Mengzhe Geng, Shoukang Hu, Xurong Xie, Mingyu Cui, Jianwei Yu, Xunying Liu, and Helen Meng. Recent progress in the cuhk dysarthric speech recognition system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2267–2281, 2021.

[26] Shailaja Yadav, Dinkar Manik Yadav, and Kamalakar Ravindra Desai. A comprehensive survey of automatic dysarthric speech recognition. *Int J Inf & Commun Technol ISSN*, 2252(8776):8776.

[27] Victoria Young and Alex Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010. PMID: 20698428.

[28] Feifei Xiong, Jon Barker, Zhengjun Yue, and Heidi Christensen. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7424–7428, 2020.

[29] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[30] Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In *Proc. Interspeech 2021*, pages 4778–4782, 2021.

[31] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Proc. Interspeech 2019*, pages 784–788, 2019.

[32] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, et al. Personalizing asr for dysarthric and accented speech with limited data. *arXiv preprint arXiv:1907.13511*, 2019.

[33] Siddharth Rathod, Monil Charola, and Hemant A Patil. Transfer learning using whisper for dysarthric automatic speech recognition. In *International Conference on Speech and Computer*, pages 579–589. Springer, 2023.

[34] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. In *Interspeech*, pages 471–475, 2018.

[35] Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. Investigation of Data Augmentation Techniques for Disordered Speech Recognition. In *Proc. Interspeech 2020*, pages 696–700, 2020.

[36] TA Mariya Celin, P Vijayalakshmi, and T Nagarajan. Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition. *Circuits, Systems, and Signal Processing*, 42(1):601–622, 2023.

[37] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[38] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[39] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[40] Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. *arXiv preprint arXiv:2401.00662*, 2024.

[41] Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukáš Burget, and Jan "Honza" Černocký. Speaker adaptation for wav2vec2 based dysarthric asr, 2022.

[42] Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen Meng. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[43] Chongchong Yu, Xiaosu Su, and Zhaopeng Qian. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1912–1921, 2023.

[44] Pu Wang and Hugo Van hamme. Benefits of pre-trained mono- and cross-lingual speech representations for spoken language understanding of dutch dysarthric speech. *EURASIP J. Audio Speech Music Process.*, 2023(1), apr 2023.

[45] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Interspeech*, volume 2008, pages 1741–1744, 2008.

[46] Linda Mortensen, Antje S Meyer, and Glyn W Humphreys. Age-related effects on speech production: A review. *Language and Cognitive Processes*, 21(1-3):238–290, 2006.

[47] Peter Torre III and Jessica A Barlow. Age-related changes in acoustic characteristics of adult speech. *Journal of communication disorders*, 42(5):324–333, 2009.

[48] Antje S. Meyer Linda Mortensen and Glyn W. Humphreys. Age-related effects on speech production: A review. *Language and Cognitive Processes*, 21(1-3):238–290, 2006.

[49] Victoria Young and Alex Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010.

[50] Wojciech A Warmbier, Małgorzata Popiel, Agnieszka Guzik, Mariusz Drużbicki, and Halina Bartosik-Psujek. Objective assessment of dysarthric disorders in patients with multiple sclerosis depending on sex, age, and type of text read. *Frontiers in Neurology*, 14:1225754, 2023.

[51] 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, 2015.

[52] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46:523–541, 2012.

[53] 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.

[54] Kaitlin L Lansford, Stephanie A Borrie, and Tyson S Barrett. Regularity matters: Unpredictable speech degradation inhibits adaptation to dysarthric speech. *Journal of Speech, Language, and Hearing Research*, 62(12):4282–4290, 2019.

[55] E Jeffrey Metter and Wayne R Hanson. Clinical and acoustical variability in hypokinetic dysarthria. *Journal of communication disorders*, 19(5):347–366, 1986.

[56] Christina Kuo and Kris Tjaden. Acoustic variation during passage reading for speakers with dysarthria and healthy controls. *Journal of communication disorders*, 62:30–44, 2016.

[57] Rupal Patel. Prosodic control in severe dysarthria. 2002.

Appendices

A Loss and WER Dynamics

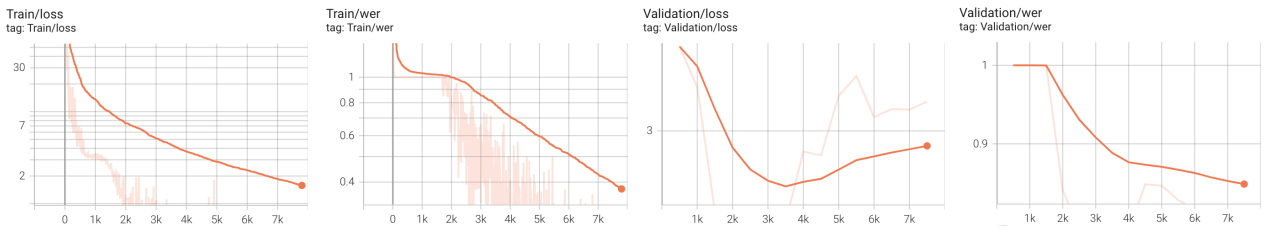

Figure 2: Model 1: Fine-tuned on dysarthric speech only

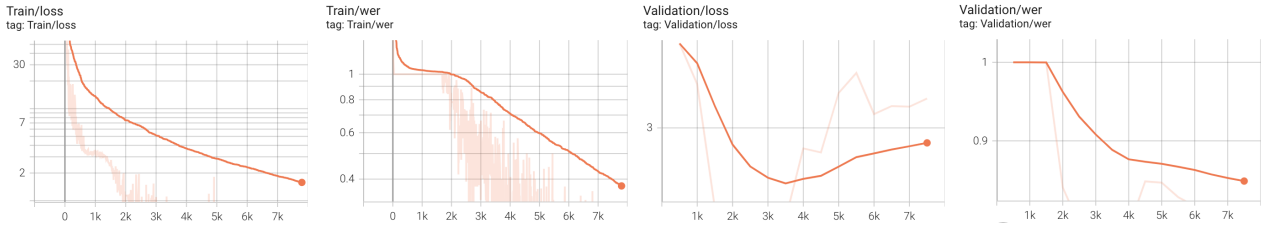

Figure 3: Model 2: Fine-tuned on dysarthric and age-matched healthy speech

Figure 4: Model 3: Fine-tuned on dysarthric and plain healthy speech

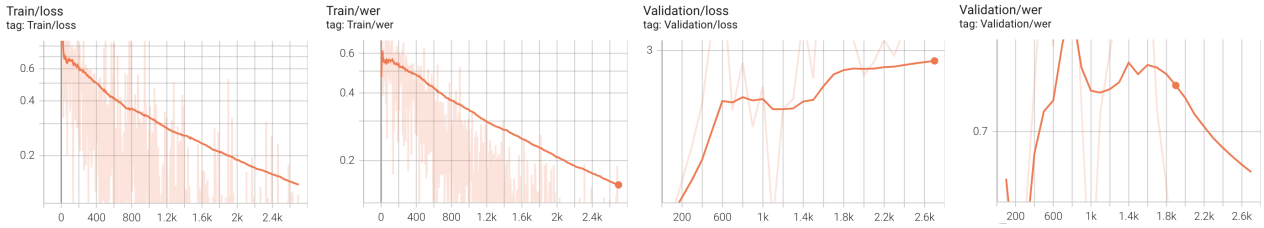

Figure 5: Model 4: Re-fine-tuned on dysarthric and plain healthy speech for high intelligibility speaker

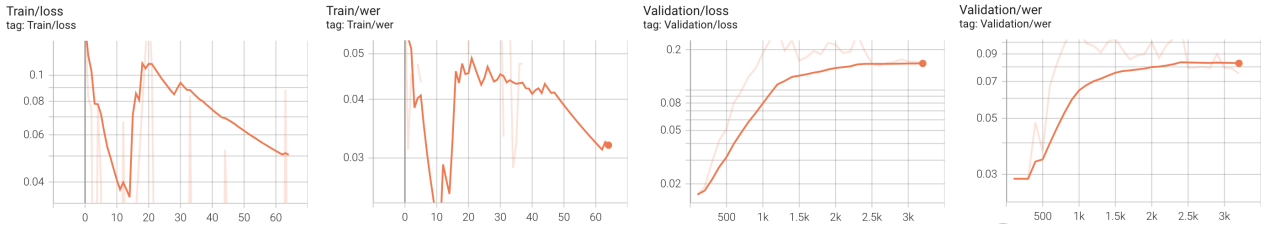

Figure 6: Model 5: Re-fine-tuned on dysarthric and age-matched healthy speech for high intelligibility speaker

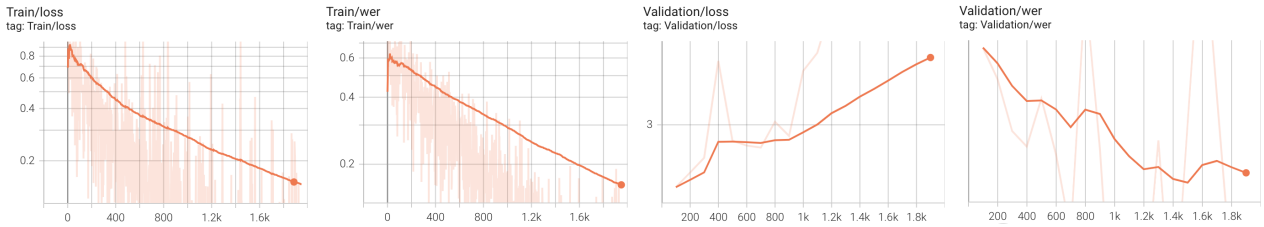

Figure 7: Model 6: Re-fine-tuned on dysarthric and plain healthy speech for low intelligibility speaker

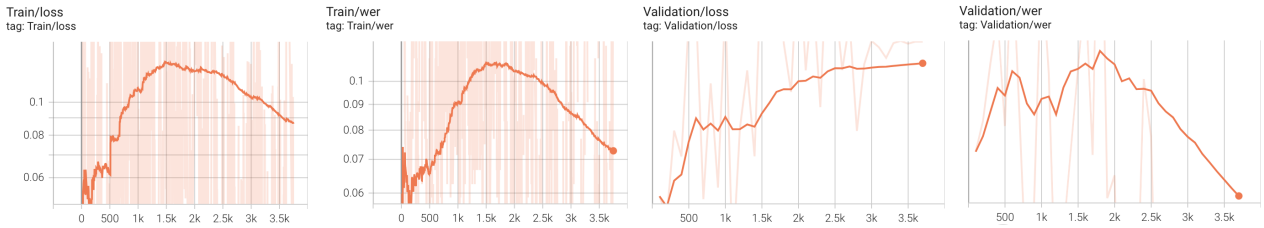

Figure 8: Model 7: Re-fine-tuned on dysarthric and age-matched healthy speech for low intelligibility speaker