university of groningen

campus fryslân

A Cross-Lingual Approach to Dutch Dysarthric Speech Recognition

Amber Lankheet

June 2025

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

A Cross-Lingual Approach to Dutch Dysarthric Speech Recognition

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
Dr. Vass Verkhodanova (Voice Technology, University of Groningen)
with the second reader being
Phat Do (Voice Technology, University of Groningen)

Amber Lankheet (S4367073)

June 11, 2025

Acknowledgements

I would like to thank my supervisor, Vass Verkhodanova, for her constant support, helpful feedback, and the brainstorming sessions that helped in shaping this thesis. Her guidance really improved the quality of my work.

I also would like to thank Phat Do for his technical advice and for helping when I ran into problems. I also want to thank Spyretta Leivaditi for answering my many questions, and for sharing the detailed GitLab repository that gave me a strong starting point for this project.

I am grateful to my boyfriend Koen Markerink for his steady support, he always believed in me and a big thanks for the great meals he made during this time. I also want to thank my friend Caz Saaltink for his programming help and encouragement when things got difficult. I also want to thank my parents, who always believed in me and supported me in all my decisions I made throughout my time as a student.

Also, I am very thankful to the Voice Technology team for being so kind and flexible. Taking a break and then coming back felt smooth and welcoming because of your support. The new students also made me feel accepted and never judged, and I really appreciate that.

Finally, I want to thank the Center for Information Technology at the University of Groningen for their technical help and for giving me access to the Hábrók high-performance computing cluster, which was essential for running my experiments.

Abstract

This thesis explores how well a multilingual self-supervised speech recognition model, XLSR-53, can understand Dutch speech from people with dysarthria, a motor speech disorder that affects pronunciation. Automatic Speech Recognition (ASR) can help people with dysarthria communicate more easily, but current systems often fail because of unclear or unusual speech patterns. A common idea in recent research is that using data from many languages (cross-lingual training) might help models better handle this kind of variation.

To test this, I compared four setups: using a high-resource Dutch model without extra training, fine-tuning on healthy Dutch speech, fine-tuning on English dysarthric speech, and a combination of both. I evaluated each model's performance using Word Error Rate (WER) on Dutch dysarthric test data. Although none of the fine-tuned models outperformed the high-resource baseline, the combined approach did slightly better than the models fine-tuned on only one type of data.

The findings show that fine-tuning with mismatched or limited data can make performance worse, even when using advanced models. This research gives insight into what does and doesn't work for dysarthric speech recognition. It also highlights important issues, such as limited speaker diversity and age differences in the data, and suggests future research could focus on phoneme-level evaluation and training specific parts of the model to improve results.

Overall, this work helps researchers and developers better understand how to create more inclusive ASR systems that support people with speech impairments.

Contents

1 Introduction

A speech disorder can make the simplest conversation a challenge, since effective communication involves multiple interdependent processes. For this reason, even basic conversations can become significantly more difficult for individuals with a speech disorder (Hernandez et al., 2022). As more people age globally, neurological disorders are becoming more common (World Health Organization, 2024). Dysarthria is an acquired or developmental speech disorder caused by neuromuscular disturbances that affect the articulators that are involved in articulation, such as jaws and tongue. It may also arise secondary to neurological diseases such as Parkinson's, Alzheimer's and Traumatic Brain Injury (TBI) (Darley, Aronson, & Brown, 1969; Joy & Umesh, 2018; Young & Mihailidis, 2010). In addition, individuals with dysarthria often experience other motor impairments, such as those caused by a TBI, which can make the use of keyboards and phones particularly challenging (Hux, Rankin-Erickson, Manasse, & Lauritzen, 2000; P. Wang & Van Hamme, 2023). Because dysarthria interferes with effective communication, it is important that these individuals have access to alternative ways of using technology. This speech impairment can lead to difficulties in social interactions with family or friends, and may also create barriers in academic or professional settings. In this context, Automatic Speech Recognition (ASR) has become increasingly important in daily life. People use smart devices and virtual assistants to perform a wide range of tasks. These hands-free technologies are especially beneficial for individuals with physical or neuro-motor disabilities who may be unable to use standard input methods like a computer mouse or keyboard (Jaddoh, Loizides, & Rana, 2023; Young & Mihailidis, 2010).

In addition, speech recognizers, as well as speech synthesizers, play a major role for individuals with speech disorders (Doyle et al., 1997; Ferrier, Shane, Ballard, Carpenter, & Benoit, 1995). ASR systems can improve the interaction capability, assist in therapy (Vaquero et al., 2006), and speed up working with a computer since there is no need for typing, according to a case study by Hux et al. (2000). The authors asked their participant, who survived a severe Traumatic Brain Injury (TBI) which caused dysarthria, to type on a keyboard and the average typing speed was 10 words per minute. Using ASR instead of the keyboard significantly reduces the physical effort required for communication or computer use and enables faster, more natural interaction. In this case, ASR could support the user in completing work tasks more efficiently and with less fatigue. This example highlights how improving ASR for dysarthric speakers is not just a technical goal. it can directly impact users' daily productivity, autonomy, and participation in work or study environments.

An example of a potential ASR application is a Personal Emergency Response System (PERS). Older adults, 65 years or older, and people with neuro-motor disabilities are at higher risk during emergencies due to limited mobility. Traditional PERS devices, often installed in homes, offer 24-hour emergency access but typically rely on panic buttons. These are sometimes avoided due to cost, stigma, or physical difficulty, and can result in false alarms. Speech-based PERS could offer a more user-friendly solution by eliminating the button, reducing stigma, and allowing false alarm cancellation (Young & Mihailidis, 2010). Such systems also support aging at home and reduce healthcare costs—every dollar spent on PERS has been linked to $7.19 in healthcare savings (Mann, Belchior, Tomita, & Kemp, 2005). However, ASR in this context must handle challenges like aging voices, stress, and speech impairments. Simpler systems with easy words and minimal training may improve reliability (Young & Mihailidis, 2010).

However, even with a simpler system, speech recognition accuracy is consistently and significantly lower for individuals with moderate to severe dysarthria compared to individuals without dysarthria P. Wang and Van Hamme (2023); Young and Mihailidis (2010). The problems for these recognition systems are disfluencies, inconsistencies and variations in speech articulation, caused by the lack of coordination in the muscles used for speaking. Specifically, irregular phoneme articulation such as imprecise production of consonants and distortion of vowels, monotone nature in loudness and pitch, slow speaking rate, slurring and mumbling. Although their speech is syntactically correct, their difficulty lies in the correct pronunciation of words, with their speech being unintelligible (Darley et al., 1969; Joy & Umesh, 2018). Mengistu and Rudzicz (2011) found that 83% of errors in ASR for a dysarthric dataset, which consists of sentences as well as words, were single-word utterances. Examples for mistakes made by people with dysarthria are consonant cluster reductions; *play* becomes [peɪ] or initial /s/ deletion; *spark* becomes [pɑrk]. A more expanded overview of these characteristics can be found in Section 2.2. These mistakes vary between individuals with dysarthria, which makes it difficult for ASR generalize over all these different dysarthric pronunciations. This results in a model that adapts more to one speaker instead of a model that can be used by everyone with dysarthria, which I will highlight more in Section 2.3.1.

ASR systems for individuals with dysarthria have shown some progress over time. STARDUST is a system by Parker, Cunningham, Enderby, Hawley, and Green (2006), that was developed to improve recognizing severe dysarthric speech based on an Hidden Markov Model (HMM). It uses data from speakers with and without dysarthria that had to articulate 10 words several times. After training on dysarthric speech, the model improved with 5% recognition accuracy compared to their baseline without the training. A more recent development, Joy and Umesh (2018) improved ASR for dysarthric speech by using a Gaussian Mixture Model and Deep Neural Network based Hidden Markov Model. The amount of utterances was a lot more than for the STARDUST project and here the authors included the TORGO dataset, which also has full sentences. It improved the Word Error Rate (WER) from 46.22% to 28.60%, which means that the lower the percentage of WER, words are recognized correctly. This WER is still high compared to typical speech, which has a WER benchmark of 4.8% for Wav2Vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020).

One way to adapt speech recognition models to dysarthric speech is by using accented English combined with dysarthric English, as explored by Shor et al. (2019). Most existing models, however, are fine-tuned and tested within the same language or are not fine-tuned at all (Hernandez et al., 2022; Shor et al., 2019; P. Wang & Van Hamme, 2023), limiting their ability to generalize across languages or speech types. To address this, cross-lingual models such as Wav2Vec-XLSR-53 may offer improvements in recognizing dysarthric speech, as these models are exposed to a wide variety of phonemes across languages. For example, in English, it is common to replace /t/ with a glottal stop, as in *sort of* pronounced [sɔʔ ɔf]. A similar pattern appears in Dutch dysarthric speech, where *wakker* (awake) may be produced as [ʋɑʔər] instead of [ʋɑkər]. These examples show that a phenomenon considered typical in one language might be interpreted as impaired speech in another (Rietveld & Van Heuven, 2016), highlighting the value of cross-lingual models in capturing such variations.

These cross-linguistic similarities in phoneme variation highlight the potential of multilingual models to generalize across both typical and impaired speech patterns. In this context, the scarcity of open-source dysarthric data—often limited due to privacy concerns and participant fatigue—further

increases the value of cross-lingual training strategies (Jaddoh et al., 2023). To explore this, Hernandez et al. (2022) evaluated self-supervised models that were pre-trained on large unlabeled corpora to learn speech representations from raw audio, either in monolingual or multilingual settings. Their study tested the models on English dysarthric speech, comparing Wav2Vec 2.0 and HuBERT (both pre-trained on English) with XLSR-53, which was trained on data from 53 languages. XLSR-53 outperformed the others, achieving a WER of 26.1%, suggesting that exposure to a wider range of linguistic input enhances model robustness.

This study investigates whether cross-lingual fine-tuning with dysarthric speech data is necessary to improve recognition performance, or if monolingual fine-tuning with typical speech is sufficient. Using a self-supervised model pre-trained on 53 languages, the model will be fine-tuned on English dysarthric speech and tested on Dutch dysarthric speech to evaluate cross-lingual generalization. While previous work has shown that self-supervised models can improve accuracy when fine-tuned with dysarthric speech from another language (Javanmardi, Kadiri, & Alku, 2024), a key research gap is the limited availability of open dysarthric datasets. This often results in speaker-specific models with poor generalizability. This study builds on findings that multilingual pretraining improves performance, which has been done by Hernandez et al. (2022); P. Wang and Van Hamme (2023), and to discover if cross-lingual fine-tuning can further reduce word error rates (Javanmardi et al., 2024). By exploring cross-lingual fine-tuning, this research aims to advance dysarthric speech recognition and address the limitations of current models.

1.1 Research Questions and Hypotheses

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

> **Does cross-lingual fine-tuning with English dysarthric speech, instead of monolingual fine-tuning with healthy speech, improve the performance of the self-supervised model XLSR-53 for Dutch dysarthric speech in ASR?**

This main question can be broken down into the following sub-questions:

- How does a high-resource checkpoint of XLSR-53 perform when tested on Dutch dysarthric speech?
- How does the performance of XLSR-53 on Dutch dysarthric speech compares to the baseline when fine-tuned on Dutch typical speech?

Based on prior work, I hypothesize that fine-tuning XLSR-53, a self-supervised, cross-lingual model, on dysarthric speech will yield a 10% reduction in Word Error Rate (WER) relative to monolingual fine-tuning. This expectation is grounded in an 8.5% WER improvement reported when fine-tuning on accented speech, where phonetic deviations such as altered vowel quality and non-standard consonant articulation mirror key features of dysarthric speech. These shared characteristics help train models to better handle pronunciation variability (Shor et al., 2019). XLSR-53 was pre-trained on 53 languages and later fine-tuned using 32 hours of high-quality speech for ASR tasks. This robust multilingual and high-resource foundation enables the model to generalize well across languages and speaker types. By exposing the model to a wide range of phonetic patterns across languages, cross-lingual fine-tuning improves its ability to cope with pronunciation variability. An additional fine-

tuning stage of dysarthric English speech should reinforce shared acoustic representations, benefiting recognition in Dutch dysarthric speech as well (Hernandez et al., 2022; P. Wang & Van Hamme, 2023).

Moreover, Shor et al. (2019) demonstrated that even a few minutes of dysarthric speech can significantly improve WER: 71% of the relative WER gain was achieved using only 5 minutes of training data. This supports the plausibility of achieving meaningful improvements with limited dysarthric speech. Therefore, a 10% WER reduction is a conservative yet realistic goal, especially considering that current cross-lingual systems still struggle with severe dysarthria, achieving only 60% recognition accuracy in such cases (P. Wang & Van Hamme, 2023).

To validate these hypotheses, I conduct three experiments. The first is a high-resource Dutch baseline using an XLSR-53 checkpoint that has already been fine-tuned on approximately 32 hours of Dutch speech, which is the only available "pre-fine-tuned" model. The second is a duration-matched Dutch condition, where that same checkpoint is fine-tuned on 140 minutes of healthy Dutch speech to allow a fair comparison with the dysarthric data. Finally, the third is a cross-lingual dysarthric condition, where XLSR-53 is fine-tuned on 140 minutes of English dysarthric speech from the TORGO corpus. These three setups enable a direct comparison between monolingual and cross-lingual fine-tuning, as well as between high-resource and low-resource training scenarios.

1.2 Thesis Outline

The structure of this thesis is as follows: The Introduction section 1 outlines the background and highlights the research gaps addressed in this study. It also introduces the research question, hypotheses, and the overall structure of the thesis. The Literature Review section 2 describes the search strategy and provides an overview of the relevant papers. It discusses existing research on dysarthric speech and datasets, ASR for dysarthric speech, and the use of cross-lingual fine-tuning. Special attention is given to speaker-dependent models and self-supervised learning. The Methodology section 3 details the datasets used for fine-tuning and evaluation. It also explains XLSR-53, the evaluation procedure, and the statistical analyses. Ethical and privacy considerations, as well as bias and fairness, are addressed in this section. The Experimental Setup section 4 covers the technical setup, including data splitting, experimental design, and hyperparameter configuration. The Results section 5 presents the outcomes of the experiments, comparing ASR performance across different fine-tuning strategies, supported by statistical tests. The Discussion section 6 interprets the experimental findings and considers possible explanations for the results. It also discusses the study's limitations. Finally, the Conclusion section 7 summarises the thesis, outlines the main findings, and provides suggestions for future research as well as the potential impact and relevance of the work.

2 Literature Review

In this section of my thesis I will provide a review of the literature existing on dysarthric speech recognition and self-supervised cross-lingual models to improve dysarthric ASR. In Section 2.1, I will explaining my search strategy and the selection of certain criteria to make this literature review replicable. Then, in Section 2.2 I will discuss what dysarthric speech is and its characteristics, with a follow-up of different dysarthric speech datasets used in research. In Section 2.3, I will discuss dysarthric speech recognition in the past, which will lead to a discussion of current research and some different approaches of creating speaker-dependent models and self-supervised learning. Finally, in Section 2.4 I will close this chapter with discussing cross-lingual fine-tuning for dysarthria and previous research about this topic.

2.1 Search Strategy and Selection Criteria

To make the literature review replicable, Google Scholar was used. The XLSR-53 model I used is retrieved from Hugging Face[1] and Github[2].The search terms were:

- (asr OR automatic speech recognition OR voice assistant) AND (dysarthric speech OR dysarthria); after 2000
- (torgo dataset OR dysarthria); after 1969
- (cross-lingual AND dysarthria); after 2020
- (xlsr-53 AND fine-tuning AND dysarthria); after 2020

Reference	**Description**
Young and Mihailidis (2010)	Challenges using ASR for dysarthric and elderly speakers
Jaddoh et al. (2023)	How people with dysarthria interact with ASR systems
Hux et al. (2000)	Systems performed better for a speaker without dysarthria than for one with mild dysarthria
Parker et al. (2006)	STARDUST project developed speaker-dependent ASR
Shor et al. (2019)	Presents finetuning techniques to improve ASR for users with ALS and accented speech
Qian and Xiao (2023)	The rise of deep learning methods since the 2010s

Table 1: Results First Search Entry

[1] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-dutch
[2] https://github.com/facebookresearch/fairseq

The literature search for this study was carried out on Google Scholar using the entry (asr AND dysarthria). To focus on the most relevant developments in automatic speech recognition, only articles published after the year 2000 were considered. In addition, all selected studies were peer-reviewed to ensure academic quality. Special attention was given to literature reviews, as these helped frame the challenges of dysarthric speech recognition and highlighted existing approaches within the field. The search returned around 16,700 results. To manage scope, only the first two pages of results were reviewed, as Google Scholar ranks articles primarily by relevance, considering factors such as keyword match and citation frequency. Studies with a clinical or therapeutic focus on dysarthria were excluded, as were articles published before 2000. The aim was to concentrate on speech technology applications rather than medical or diagnostic research. Based on these criteria, the selected articles included in this literature review are listed in Table 1.

Reference	**Description**
Rudzicz, Namasivayam, and Wolff (2012)	Information about TORGO database
Joy and Umesh (2018)	Presents improved DNN-HMM ASR models for dysarthric speech in the TORGO dataset
Darley et al. (1969)	Diagnostic patterns of dysarthric speech
Mengistu and Rudzicz (2011)	Speech characteristics of dysarthric speech
Van Nuffelen, De Bodt, Middag, and Martens (2009)	Dutch corpus of pathological and normal speech (COPAS)
Menendez-Pidal, Polikoff, Peters, Leonzio, and Bunnell (1996)	Nemours corpus
H. Kim et al. (2008)	UASpeech corpus
M. Kim, Kim, Yoo, Wang, and Kim (2017)	Korean dysarthric speech without severe dysarthria
Ons, Gemmeke, and hamme (2014)	Domotica dataset

Table 2: Results Second Search Entry

For the second literature entry, the search was conducted using the keywords (torgo dataset OR copas OR dysarthria), with a publication date filter starting from 1969. This search resulted over 8,000 results, from which the first three pages on Google Scholar were reviewed. The focus of this search was on studies that address dysarthric speech data, particularly the variability in pronunciation among individuals with dysarthria. Included papers were required to be peer-reviewed and could also include clinical research related to the causes of dysarthria, as these help contextualize speech variation within the datasets. In contrast, studies primarily concerned with assessing the severity of dysarthria were excluded, as they fall outside the scope of this research. The selected studies meeting these criteria are presented in Table 2.

For the third search entry (cross-lingual AND dysarthria), I focused on papers published from 2020 onwards, since I mostly focused on self-supervised model XLSR-53, which was introduced in that

Reference	Description
Hernandez et al. (2022)	Self-supervised models like Wav2Vec, HuBERT, and XLSR improve ASR performance on dysarthric speech
P. Wang and Van Hamme (2023)	Different pre-training strategies for spoken language understanding (SLU) systems on Dutch dysarthric speech
Baevski et al. (2020)	Self-supervised model wav2vec 2.0
Javanmardi et al. (2024)	Improved generalization and accuracy after fine-tuning in cross-database scenarios
Conneau, Baevski, Collobert, Mohamed, and Auli (2020)	Introduces XLSR, a cross-lingual speech representation model

Table 3: Results Third Search Entry

year. The search gave 389 results, and I looked through the first 3 pages on Google Scholar to find relevant studies. Only peer-reviewed papers were included to make sure the sources are reliable. I chose not to include studies that focus on classifying the severity of dysarthria, because this thesis is not about identifying how severe someone's speech disorder is. Instead, it focuses on improving speech recognition for all types of dysarthric speech. Papers that met these criteria are listed in Table 3.

For the final search entry (xlsr-53 AND fine-tuning AND dysarthria), I again limited the results to papers published after 2020, since XLSR-53 was introduced that year. This search term is closely related to the previous one, so many of the same results appeared. Google Scholar returned 48 articles in total, and I reviewed the first 2 pages to identify relevant studies. In addition to direct search results, some key papers were found through references in other studies. For example, the papers by Rosen and Yampolsky (2000) and Mann et al. (2005) were cited in Young and Mihailidis (2010), while Espana-Bonet and Fonollosa (2016) was found through Joy and Umesh (2018). The papers on different self-supervised learning models (Chen et al., 2022; Graham & Roll, 2024; Hsu et al., 2021) were referenced in Su (2024), which influenced the research direction taken in this thesis.

2.2 Dysarthric Speech

In this section I will get in depth about the cause of dysarthria and the characteristics of dysarthric speech. Dysarthria is a motor speech disorder caused by damage to the central or peripheral nervous system, leading to weakness or incoordination of the muscles involved in speech. It affects the physical execution of speech rather than language processing or speech planning. The disorder can affect several aspects of speech. Respiration may be impaired, resulting in reduced breath support and short, quiet utterances. Articulation can become slurred or imprecise, while poor control of the velopharyngeal mechanism can lead to hypernasal resonance. Prosody may also be disrupted, leading to speech that sounds monotone or lacks natural rhythm. These features vary depending on the type of dysarthria, such as flaccid, spastic, ataxic, hypokinetic, or hyperkinetic dysarthria,

with each type reflecting a distinct pattern of neuromotor impairment as described by Darley et al. (1969). Importantly, dysarthric speech is highly individual. Not every characteristic appears in every speaker, and the severity and combination of symptoms differ across cases.

2.2.1 Dysarthric Speech Characteristics in TORGO

The TORGO[3] English database is designed to support the development of speech recognition systems for people with dysarthria, particularly spastic, ataxic, and athetoid types caused by cerebral palsy, as well as one case of ALS. It includes recordings from six dysarthric and matched control speakers, along with detailed articulatory data. Their age range is between 16 and 50 years old. The speech material spans non-words, short words, structured sentences, and spontaneous speech, enabling analysis of a wide range of phonetic contrasts. Observed speech deviations are grouped into different classes, making the dataset valuable for both ASR research and clinical studies (Rudzicz et al., 2012). In Section2.2.2, I will discuss other dysarthric datasets as well. The observed deviations in the TORGO dataset of six subjects are grouped into different classes (Mengistu & Rudzicz, 2011):

- **Final consonant deletion:** Omission of the final consonant, which requires more articulatory control. E.g: Feed $\rightarrow$ Fee
- **Consonant cluster reduction:** Omission of a consonant in a consonant cluster. E.g: Grow $\rightarrow$ Gow
- **Initial /s/ deletion:** When the /s/ is followed by a stop. E.g: Spark $\rightarrow$ Park
- **Devoicing:** The voiceless counterpart is pronounced E.g: Deer $\rightarrow$ Teer
- **Fronting:** Consonants that are normally produced at the back of the alveolar ridge are substituted by consonants that are produced at or in front of the alveolar ridge. E.g: Ship $\rightarrow$ Sip
- **Vocalization:** Liquids (/l/ and /r/) are sometimes produced as vowels when they occur in word-final positions. E.g: Table $\rightarrow$ Tabo
- **Stopping:** Substitution of a stop consonant for a fricative. E.g: Thorn $\rightarrow$ Torn

To analyze pronunciation differences, the expected English phoneme sequences were compared to the actual phoneme sequences produced by the dysarthric speakers. The expected English phonemes are based on the CMU Pronunciation Dictionary[4]. This is an open-source machine-readable pronunciation dictionary for North American English. Other observed articulatory mistakes were poor articulation of vowels. However, these characteristics are specific to the TORGO dataset, which contains English dysarthric speech recorded from speakers in Toronto. While dysarthria presents core features across languages, such as imprecise articulation, reduced intelligibility, and abnormal prosody, its manifestation is shaped by each language's phonetic structure. In English, reduced vowel clarity and complex consonant clusters, such as /str/, /pl/, are particularly affected. In contrast, Dutch dysarthric speakers often struggle with the production of uvular fricatives like /χ/ and

[3]https://www.cs.toronto.edu/ complingweb/data/TORGO/torgo.html

[4]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

/ɣ/, as well as uvular /r/, which may be omitted or replaced with a glottal or alveolar variant. Voicing contrasts in Dutch, such as /p/ vs. /b/, can also be weakened, leading to mergers that reduce intelligibility (P. Wang & Van Hamme, 2023).

Some individuals with dysarthria tested in the TORGO dataset differed in these neurological disorders, which adds context to variability between the speakers (Rudzicz et al., 2012). This variability arises not only from the wide range of dysarthria types and severities but also from individual differences in speech patterns, articulation strategies, and compensation techniques. Moreover, collecting high-quality speech data from individuals with dysarthria is challenging. Participants often find the recording sessions physically and mentally demanding, especially when tasks involve repeated articulation or prolonged speech. As a result, available datasets are limited in size and scope. In the next section, I will discuss different datasets that are often used in research for dysarthric speech recognition.

2.2.2 Dysarthric Speech Datasets

Because of the data scarcity and the variability of dysarthric speech, datasets are quite limited (Jaddoh et al., 2023). Current datasets mostly consist of repetitions of utterances by the same participant and are lacking unique utterances. As I mentioned in the introduction as well, physical fatigue as well as frustration of the speaker are also reasons why the amount of data is little. The most-used datasets for research of dysarthric speech are Nemours (0.9 hours) (Menendez-Pidal et al., 1996), UASpeech (17 hours) (H. Kim et al., 2008) and TORGO (23 hours) for English (Joy & Umesh, 2018; Young & Mihailidis, 2010). The last two are currently widely used in ASR research. Due to the unavailability of UASpeech, TORGO is used for English dysarthric speech. However, healthy speech is also included in the amount of hours which makes the amount of dysarthric speech data smaller than the overall dataset size. Some datasets do not contain severe dysarthria, since the researchers could not recruit people with severe dysarthria, but only mild or moderate, for example a dataset retrieved from M. Kim et al. (2017). This data consisted of mild and moderate Korean dysarthric speech.

For Dutch, the datasets are even more limited. In P. Wang and Van Hamme (2023), the Dutch dysarthric dataset used for their research was Corpus Pathologische en Normale Spraak (COPAS)[5]. This is a Dutch corpus of pathological speech recorded in Flanders, a Dutch-speaking region in Belgium (Van Nuffelen et al., 2009). In the paper by P. Wang and Van Hamme (2023), the authors give an overview of the amount of dysarthric data, which is around 4 hours with different severities. Another Dutch dysarthric dataset discussed by P. Wang and Van Hamme (2023) is the Domotica[6] dataset (Ons et al., 2014). This is again a Flemish speech corpus designed for home automation applications, containing 4,147 utterances from 17 speakers with varying levels of speech impairment. The dataset includes commands for controlling domestic devices, such as adjusting lights, doors, and heating.

According to Jaddoh et al. (2023), the authors argue that using dysarthric speech for training data improves the performance of speech recognition. Including users with dysarthria in designing and testing ASR systems will eventually solve the problem of data scarcity and improve the performance.

[5]https://taalmaterialen.ivdnt.org/download/tstc-corpus-pathologische-en-normale-spraak-copas/
[6]https://www.esat.kuleuven.be/psi/spraak/downloads/

In contrast to that, Young and Mihailidis (2010) argues that collecting more dysarthric speech data will be problematic, since it can cause fatigue when the same individual has been used. Also, using more individuals for the collection of data can be very time consuming, since people with dysarthria are very vulnerable. Their solution for data scarcity is a simple corpus which minimizes the required training time and will benefit speech-based Personal Emergency Response Systems (PERS), as mentioned in Section 1. However, this system cannot be used for everyday communication with family or as a substitute for a mouse or keyboard during work or study.

2.3 Dysarthric Speech Recognition

Recognizing dysarthric speech comes with many challenges, such as the lack of large datasets, high variability between speakers, and frequent disfluencies. A traditional approach for addressing these challenges involves using Hidden Markov Models (HMMs) combined with Gaussian Mixture Models (GMMs) (Joy & Umesh, 2018; Mengistu & Rudzicz, 2011; Parker et al., 2006). These models are considered interpretable because their structure, such as states, transitions, and emission probabilities which can be examined and understood in a transparent way. This allows researchers to analyze how the model makes predictions and to adjust its parameters based on prior linguistic or acoustic knowledge. It is a well-understood framework because it has been widely used and tested in speech processing for decades, especially in low-resource settings. HMM-GMM systems can work relatively well with limited data, as shown in the STARDUST project (Parker et al., 2006), where speaker-dependent HMMs were used successfully with only eight participants who had severe dysarthria. Similarly, the TORGO dataset was used in earlier research on adapting these models for dysarthric speakers (Joy & Umesh, 2018; Mengistu & Rudzicz, 2011), demonstrating that even with restricted vocabulary and variable speech patterns, consistent phonetic tokens within each speaker's output could be used to train effective recognizers. Nonetheless, a GMM-HMM struggles with large-vocabulary tasks due to the speech characteristics of dysarthria. The performance of a GMM-HMM did improve over time based on Joy and Umesh (2018) and Espana-Bonet and Fonollosa (2016). The monophone GMM-HMM by Espana-Bonet and Fonollosa (2016) had a WER range of 29.10%-88.62%, where Joy and Umesh (2018) improved this with a monophone GMM-HMM with speaker-dependent transformations. Their WER range was 20.28%–80.92%, which was enhanced by tuning key parameters such as frame shift duration, context-dependent states, and feature dimensionality specifically for dysarthric speech. This demonstrates that parameter optimization is effective.

In addition to using GMM-HMM models, Joy and Umesh (2018) also explored DNN-HMM hybrids, which combine the traditional HMM framework with deep neural networks. They applied sequence-discriminative training, specifically State Minimum Bayes Risk (sMBR), and achieved a 17.62% relative reduction in WER compared to the model from Espana-Bonet and Fonollosa (2016). DNN-HMM models outperform GMM-HMMs for dysarthric speech because they can capture more complex, non-linear patterns in speech and better handle speaker variability. This has also been confirmed by recent findings in speech technology, where DNN-HMMs are widely used due to their flexibility and improved accuracy (Qian & Xiao, 2023). However, these models require large amounts of labeled data, which can be a problem in dysarthric ASR since many individuals cannot provide enough annotated speech for training. As a result, there is a trade-off: DNNs generally perform better but rely on large datasets, while GMMs work with less data but offer weaker results.

The study by Joy and Umesh (2018) focused on speaker-dependent adaptations, meaning that WER improvements varied across different severity levels of dysarthria. I will discuss this further in Section 2.3.1. Finally, neither GMMs nor DNNs are pre-trained on multilingual data, which limits their ability to generalize across languages. This highlights the potential of cross-lingual self-supervised models like XLSR-53, which are trained on many languages and require less labeled data for effective adaptation.

Building on this need for generalizability and data efficiency, recent approaches have turned to transfer learning to further improve dysarthric ASR. In particular, Joy and Umesh (2018) applied transfer learning by training a teacher DNN on dysarthric-only speech, which then guided a student DNN trained on both dysarthric and control speech. This technique transferred useful knowledge via soft targets but still relied on large amounts of labeled dysarthric data, which are difficult and time-consuming to collect. Furthermore, the method required significant computational resources, limiting its practicality. To overcome these challenges, Shor et al. (2019) proposed a lighter fine-tuning approach that required only 5-10 minutes of dysarthric speech, combined with accented English. Their model, based on a pretrained RNN-Transducer with a Listen, Attend and Spell module, showed a 70% WER improvement over a baseline trained only on healthy speech. However, since the comparison involved different architectures, including Google Cloud ASR, the results are harder to generalize. To ensure valid evaluation, this thesis uses the same model architecture for both the baseline and the fine-tuned systems.

2.3.1 Speaker-Dependent Models

As previously mentioned, dysarthric speech varies greatly between individuals speaking the same language, due to differences in severity and the location of neurological damage (Darley et al., 1969). For typical speech, speaker-independent models, trained on many speakers, can recognize thousands of words and generalize well to new users (Young & Mihailidis, 2010). However, developing ASR systems for dysarthric speech is challenging because of high variability across speakers (Jaddoh et al., 2023; Parker et al., 2006; Rosen & Yampolsky, 2000). Parker et al. (2006) found that training and testing an ASR model on the same individual, a speaker-dependent approach, can improve recognition performance for dysarthric speech. On the other hand, Young and Mihailidis (2010) argue against speaker-dependent models due to the large number of dysarthric speakers required to build generalized systems, and the long training times involved. They also highlight the increased fatigue experienced by dysarthric users during extensive data collection. To address this, they proposed a simplified system focused on recognizing very short words, words with 1-2 syllables, with a small vocabulary and minimal training, although no empirical testing was conducted to validate this approach. In contrast, Parker et al. (2006) support speaker-dependent systems with their STARDUST model, which achieves improved performance through personalized training. However, since STARDUST recognizes only about 10 words, it limits communication options and reduces ASR inclusivity for people with dysarthria who need more flexible and extensive vocabulary support.

In contrast to the previous articles, Rosen and Yampolsky (2000) preferred a speaker-adaptive model, which updates acoustic templates dynamically as the user speaks, improving dysarthric speech recognition. The system, which is built from scratch, adapts over time correctly to the user's specific speech patterns. The difference between a speaker-dependent and speaker adaptive model is that for speaker-dependent the model is trained on one user's data, whereas a speaker-adaptive model

is pre-trained on different speakers and it adjusts to a speaker's characteristics through fine-tuning. A more recent development is the personalized model by Shor et al. (2019). It begins with a large speaker-independent model as their base model, trained on healthy speech. Next, it was fine-tuned on a small amount of user-specific data, which was English dysarthric and accented speech. This makes the model not speaker-dependent, but speaker-adaptive, which does not start from scratch, as in Rosen and Yampolsky (2000). It thus provides better accuracy and is more practical for dysarthric speech recognition.

2.3.2 Self-supervised Learning

Self-supervised learning (SSL) is a recent development that has a promising approach to address challenges in dysarthric speech recognition, as previously mentioned the scarcity of dysarthric data and variability in speech. SSL models, unlike DNNs and GMMs, are pre-trained on large amounts of unlabeled data to learn speech representations and can be fine-tuned with smaller labeled datasets. These models try to predict masked parts of the audio signal (Conneau et al., 2020; Hernandez et al., 2022; Javanmardi et al., 2024). This makes SSL models more flexible than traditional DNNs, which often requires domain-specific feature engineering or complex training methods to adapt to dysarthric speech (Joy & Umesh, 2018). SSL models can learn layered patterns in speech, including both normal and unusual speech. For example, Javanmardi et al. (2024) showed that wav2vec 2.0 worked better than traditional MFCC-based systems when trained and tested on dysarthric speech. SSL models also handle noise and recording mismatches better, thanks to pre-training on diverse data. In contrast, traditional systems often rely on post-processing methods like FMLLR or MLLR to handle such variation (Joy & Umesh, 2018), which adds complexity and may not fully solve the problem of background noise, different microphones, and inconsistent recording environments.

Several examples of an SSL models are wav2vec 2.0 (Baevski et al., 2020), wav2vec 2.0 variants (XLSR-53 and XLS-R) (Babu et al., 2021; Conneau et al., 2020), WavLM (Chen et al., 2022) and HuBERT (Hsu et al., 2021). WavLM is a newer self-supervised learning (SSL) model that improves on wav2vec 2.0 by adding features like relative positional encoding, speech denoising, and multi-task training to better handle noise and speaker variation (Chen et al., 2022). While these upgrades have led to strong results on standard benchmarks, WavLM has been less tested on non-English or dysarthric speech. In contrast, wav2vec 2.0 has already shown strong performance in low-resource and dysarthric ASR tasks, even with limited labeled data (Baevski et al., 2020; Javanmardi et al., 2024). It also has multilingual versions, like XLSR-53, which are useful for underrepresented languages such as Dutch. For these reasons, wav2vec 2.0 serves as a strong and practical baseline for studying dysarthric speech recognition. This model encodes speech audio via a multi-layer convolutional neural network and uses a transformer-based context network to predict missing parts from the audio without needing hand-crafted features.

HuBERT is another SSL model that, like wav2vec 2.0, learns from raw audio without labels. It uses a clustering step to create pseudo-labels before training, which helps it learn useful speech patterns (Hsu et al., 2021). Some studies show that HuBERT can improve recognition for dysarthric speech, especially with advanced training tricks like adversarial data augmentation. However, it is less commonly used than wav2vec's XLSR-53 in multilingual or cross-lingual setups, and there are fewer pretrained models or benchmarks for Dutch. Because of this, XLSR-53 is a more practical and reliable choice for this research.

2.4 Cross-lingual Fine-tuning for Dysarthria

Another model that is advanced in speech recognition is OpenAI's Whisper[7] (Graham & Roll, 2024). It is trained on 680,000 hours of multilingual and multitask data, making it highly effective for recognizing a wide variety of languages, accents, and speech contexts. However, unlike self-supervised models like Wav2Vec 2.0, Whisper needs paired audio-text data for training, making it less suitable for adapting to dysarthric speech with limited labeled data (Fan, Shankar, & Alwan, 2024). Also, fine-tuning Whisper is more complex because of its encoder-decoder structure and large size, which needs more memory and careful setup. Since Wav2Vec 2.0 and its multilingual version XLSR-53 are easier to fine-tune on small datasets and have already shown strong results on dysarthric and low-resource speech, they are a more practical choice for this research.

In Section 2.3.2, I quickly mentioned XLS-R (Babu et al., 2021). This is a model build upon Wav2Vec 2.0. It is a powerful self-supervised speech model trained on over 400,000 hours of multilingual audio, offering broader coverage and stronger generalization than earlier models. However, XLSR-53, a smaller version trained on 56,000 hours across 53 languages, is more accessible and still highly effective—especially for cross-lingual fine-tuning. P. Wang and Van Hamme (2023) show that XLSR-53 outperforms mono-lingual models and even Whisper in tasks involving Dutch dysarthric speech. Therefore, XLSR-53 strikes a good balance between performance and practicality.

In addition, the strength of XLSR-53 in multilingual ASR comes from its large-scale training on diverse transcribed audio from the web, allowing it to generalize well across different linguistic features and typical variations in pronunciation. It relies on noisy transcriptions from the web rather than learning directly from unlabeled audio. In contrast, Wav2Vec 2.0 is a true SSL model that first pre-trains on large amounts of unlabeled raw audio using a contrastive, masked prediction objective, and is then fine-tuned on labeled data. This two-stage approach allows XLSR-53 to learn robust and generalizable acoustic representations that can adapt well to non-standard speech, such as dysarthria, even with limited labeled data (Conneau et al., 2020).

Recent studies have explored the effectiveness of cross-lingual self-supervised learning models like XLSR-53 for improving ASR performance on dysarthric speech. Hernandez et al. (2022) demonstrated that XLSR's multilingual phoneme representations effectively model dysarthric speech patterns without manual feature engineering, such as in DNN-HMM systems. This aligns with findings by P. Wang and Van Hamme (2023), who observed that SSL representations generalize better across languages and speaker variations compared to monolingual approaches.

Further supporting these advancements, Javanmardi et al. (2024) investigated the role of fine-tuning XLSR-53 in dysarthric speech detection across English and Italian. Their study revealed that fine-tuned XLSR features achieved absolute accuracy improvements of 1.46%–8.65% in cross-database scenarios, suggesting an advantage over monolingual models like Wav2Vec-BASE. Although they did not explicitly benchmark XLSR against monolingual SSL models within the same experimental setup, the observed improvements were attributed to XLSR's ability to capture shared dysarthric characteristics across languages. This was credited to XLSR's extensive pretraining on 56,000 hours of multilingual data, which I will discuss in Section 3.1.1. It enhances robustness to linguistic and acoustic variability, a finding that aligns with the observations of P. Wang and Van Hamme (2023)

[7] https://huggingface.co/openai/whisper-small

on cross-lingual adaptation.

Together, these studies highlight the potential of cross-lingual SSL models like XLSR-53 for dysarthric speech processing, particularly in addressing data scarcity. While Javanmardi et al. (2024) and P. Wang and Van Hamme (2023) both emphasize the advantages of multilingual pretraining, such as reduced reliance on labeled data and improved generalization, their focus diverges; Javanmardi et al. (2024) focus on how fine-tuning helps adapt pretrained models to work better with disordered speech, while P. Wang and Van Hamme (2023) highlight how XLSR-53 can already transfer well across languages without extra training. However, neither study directly compares multilingual and monolingual fine-tuned models under the same conditions.

This concludes the literature review. While earlier research has made progress in dysarthric speech recognition through traditional models, deep learning, and transfer learning, challenges such as limited data, speaker variability, and language mismatch still remain. Recent studies show that self-supervised models, especially cross-lingual ones like XLSR-53, offer strong potential for addressing these issues. By learning from large amounts of multilingual data, XLSR-53 can better handle variation in speech and reduce the need for extensive labeled datasets. Building on this, my research uses English dysarthric speech to fine-tune the model and Dutch dysarthric speech to test it. Although the model is not personalized to individual speakers, it can be considered speaker-adaptive because it is fine-tuned to better recognize dysarthric speech characteristics across multiple speakers. The methodology behind this fine-tuning approach will be discussed in detail in Section 3.

3 Methodology

In this section, I will discuss the methodology of my thesis. This includes the description of the dataset, core methods and models, technical framework and evaluation methodology. I will also consider ethics and research integrity at the end of this section. In Section 3.1 I highlight the exact datasets to make this study replicable. Then in Section 3.2, I will discuss the model I am using in detail and I will highlight the training details. Section 3.3 describes the evaluation method and metrics that are used in this thesis. This includes Word Error Rate and statistical analysis. At last, Section 3.4 will provide information about ethical considerations, including data privacy, FAIR principles, fairness and replicability.

3.1 Dataset Description

A variety of datasets is necessary for conducting this research. In the following subsections I will discuss the pre-train datasets for XLSR-53, fine-tune datasets and the testing datasets.

3.1.1 Pre-training datasets XLSR-53

The model XLSR-53 has been pre-trained on Common Voice, BABEL and Multilingual LibriSpeech (MLS), resulting in a pre-training corpus of 56k hours across 53 languages. This large-scale multilingual data enables robust cross-lingual representation learning (Conneau et al., 2020). All audio is sampled at 16 kHz.

CommonVoice[8]: This dataset comprises read speech in 38 languages, with 11 languages selected for pre-training: Spanish, French, Italian, Kyrgyz, Dutch, Russian, Swedish, Turkish, Tatar, Chinese, and English. The total pre-training data from CommonVoice is 1,350 hours, combining 793 hours from the 10 evaluation languages and 557 hours of English audio (Ardila et al., 2019).

BABEL[9]: This dataset consists of conversational telephone speech, primarily in Asian and African languages. For pre-training XLSR-53, 10 languages are used: Bengali, Cantonese, Georgian, Haitian, Kurmanji, Pashto, Tamil, Turkish, Tokpisin, and Vietnamese, totaling 650 hours. The dataset is notable for its balanced distribution of hours per language, which is ranging from 30 to 130 hours (Gales, Knill, Ragni, & Rath, 2014).

Multilingual LibriSpeech (MLS)[10]: Derived from read audiobooks, MLS includes 8 languages: Dutch, English, French, German, Italian, Polish, Portuguese, and Spanish. The English subset dominates with 44k hours, while the remaining 7 languages contribute approximately 6.7k hours (Pratap, Xu, Sriram, Synnaeve, & Collobert, 2020).

3.1.2 Fine-tuning

According to the paper by Conneau et al. (2020), it is necessary to fine-tune XLSR-53 before testing the model. Therefore, I used a model by Grosman (2021), which was fine-tuned with around 32

[8]https://commonvoice.mozilla.org/en/datasets
[9]https://catalog.ldc.upenn.edu/byyear
[10]https://www.openslr.org/94/

hours of Dutch data, 18 hours of Dutch from Common Voice 6.1 and 14 hours of Dutch from the CSS10 single-speaker dataset. For the next experiment, I fine-tuned my model on healthy Dutch speech using the Common Voice 13.0 "delta" subset, which includes a variety of Dutch dialects. The age range is unknown, since providing age data is voluntary. I considered using Corpus Gesproken Nederlands (CGN), but there were two main issues. First, CGN requires a license to access. Second, the CGN files are very large, which is exceeding the capacity of my computer. Although CGN is available on our university's cloud (UWP), I did not know how to download or manage such a big dataset from there. In comparison, Common Voice 13.0 (released in April 2023) was easy to grab, and XLSR-53 (developed in 2021 (Conneau et al., 2020)) had never seen that data before. For fine-tuning on English dysarthric speech, I used the TORGO corpus, as I mentioned in Section 2.2.2. It is important to compare fine-tuned monolingual and cross-lingual models with the same amount of fine-tuning data to make a valid comparison.

3.1.3 Testing

To test the XLSR-53 model on Dutch dysarthric speech, I used COPAS together with Domotica. The COPAS Corpus, aged between 50 and 80 years old, includes recordings from 319 speakers across eight categories, for example, normal, dysarthria, hearing impairment and laryngectomy (Van Nuffelen et al., 2009). Speakers performed various tasks like passage reading, articulation assessment, and storytelling. However, for this research, I only included the 2 read sentences of the dysarthric speakers, which are a total of 50 different speakers. I will explain the details about pre-processing and data preparation in Section 4.1.

The Domotica Dataset (Ons et al., 2014) contains over 3,000 Flemish Dutch utterances related to home automation, recorded from 17 dysarthric speakers (15 adults and 2 children) with conditions like multiple sclerosis, aged between 14 and 61 years old. For every experiment, I used the same dysarthric set to make a valid comparison between the fine-tuning with English dysarthric speech and Dutch typical speech is as similar as possible.

3.2 Model

In this subsection, I will discuss the reason for using the XLSR-53 model instead of other models and I will discuss the model architecture. For this thesis, the selection of a sufficient model is crucial to achieve reliable results. This process for choosing the right model was based on previous research. For the improvement of speech recognition for dysarthric speech, XLSR-53 has often been used. It has also been cross-lingual fine-tuned with Italian and tested on English (Javanmardi et al., 2024). There are several expanded models as I mentioned in Section 2.4, such as XLS-R and its variations. XLSR-53 and XLS-R are both cross-lingually pretrained wav2vec 2.0 models. XLS-R is using over 372K hours of unlabeled audio from 128 languages, making it more robust and broader in linguistic coverage. However, for cross-lingual fine-tuning on dysarthric speech, XLSR-53 is still sufficiently powerful. Since dysarthric datasets are small and often low-resource, XLSR-53's ability to generalize from many languages makes it a strong and more computationally efficient choice without requiring the massive scale of XLS-R (Babu et al., 2021; Hernandez et al., 2022; P. Wang & Van Hamme, 2023). Therefore, XLSR-53 is a reasonable model.

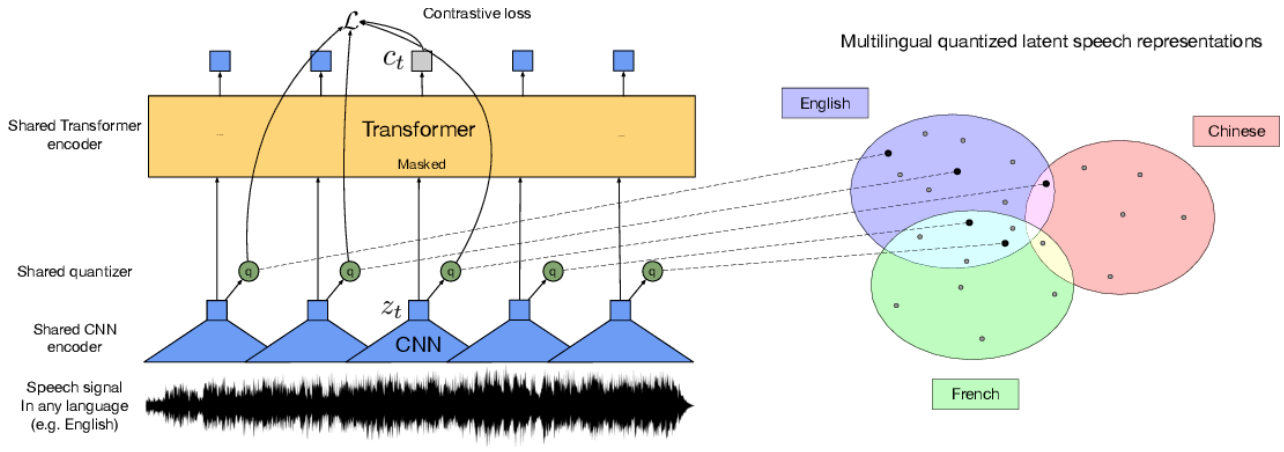


Figure 1: System architecture from XLSR-53. Reprinted from Conneau et al. (2020)

3.2.1 Model Architecture

As I previously mentioned, XLSR-53 is a model that is based on the wav2vec 2.0 framework. In Figure 1, the architecture of wav2vec-XLSR-53 is visualized, together with the multilingual quantized latent speech representations. Firstly, wav2vec2.0 is introduced by Baevski et al. (2020) and it is a self-supervised model for learning speech representations directly from raw audio. It consists of three main components: a convolutional feature encoder, a Transformer network and a quantization module.

Convolutional feature encoder. The convolutional feature encoder processes raw audio waveforms into latent representations. The CNN in Figure 1 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. The convolutional layers are followed by a layer normalization and a Gaussian Error Linear Unit (GELU) activation function, ensuring that the transformation is both non-linear and normalized. The raw audio waveform is first normalized to zero mean and unit variance, and the encoder's total stride determines the number of time steps T passed to the Transformer.

Transformer architecture. The output of the feature encoder is fed into a context network that follows the Transformer architecture, which uses self-attention mechanisms to generate contextualized representations $C = (c_1, c_2, \ldots, c_T)$. This enables the model to capture dependencies across the entire sequence of latent representations, improving its understanding of speech context. Instead of using fixed positional embeddings to encode absolute position, the model employs a convolutional layer as a relative positional embedding. The output of this convolution is passed through a GELU activation, added to the input sequence, and then layer-normalized, supporting stable and effective learning.

Quantization module. For self-supervised training, the output from the feature encoder is discretised into a finite set of speech representations using product quantization. This involves selecting quantized vectors from multiple codebooks, each containing a fixed number of entries. One entry is chosen from each of the codebooks, the resulting vectors are concatenated, and a linear transformation is applied to produce the final quantized vector. The selection process is made differentiable using the Gumbel softmax function. This approach has proven effective in previous work, enabling

the model to jointly learn discrete speech units and contextualized representations.

Building on this, XLSR-53 by Conneau et al. (2020) extends wav2vec 2.0 to the multilingual setting by using the same architecture as above, with a shared feature encoder, context network, and quantizer. The model is trained on unlabeled speech from 53 different languages. A key innovation was that the quantized speech units q_t are shared across all languages, which enables effective cross-lingual representation learning. This design helps the model to generalize better to low-resource languages (Bălan, 2023) and it will adapt to unseen language conditions, while retaining architectural simplicity. As a result, XLSR-53 learns robust speech patterns from a large, multilingual dataset and can be fine-tuned for various downstream tasks across different languages and speech impairments.

3.2.2 Training Details

Masking. As I mentioned in Section 3.2.1, XLSR-53 begins by processing raw audio through a multi-layer convolutional feature encoder, which outputs latent speech representations. During pre-training, a proportion of these latent representations are masked. Specifically, a fraction $p = 0.065$ of time steps are randomly selected as starting indices, and the subsequent $M = 10$ consecutive time steps are masked. These masked spans are replaced with a learned feature vector shared across all masked positions. This masking strategy is inspired by masked language modeling in BERT (H. Wang et al., 2024) and ensures the model learns robust contextual representations by predicting the masked segments.

Training. The pre-training objective involves a contrastive task where the model must distinguish the true quantized latent representation q_t for a masked time step from a set of distractors. The contrastive loss is defined as:

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

where c_t is the Transformer output for the masked time step, Q_t contains the true latent and K distractors sampled from other masked time steps, and $sim(a,b)$ computes cosine similarity between vectors a and b. The loss encourages the model to align the contextual representation with the correct quantized latent. Additionally, a diversity loss L_d maximizes the entropy of codebook usage to ensure all entries are utilized:

$$L_d = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\bar{p}_{g,v}log(\bar{p}_{g,v})$$

where $\bar{p}_v, g$ is the average softmax probability for codebook entry v in group g.

Fine-Tuning. After pre-training, the model is fine-tuned on labeled data. A linear projection layer is added on top of the Transformer to map contextual representations to output tokens. The model is trained using Connectionist Temporal Classification (CTC) loss, which aligns audio sequences with transcriptions without requiring explicit segmentation. During fine-tuning, SpecAugment is applied to mask time steps and channels, enhancing robustness. The feature encoder remains frozen, while the Transformer and output layer are updated.

3.3 Evaluation Methodology

In this subsection, I will discuss the evaluation method that I used for evaluating speech recognition for dysarthric speech and I will discuss statistical tests to evaluate the significance of the experiments in this thesis.

3.3.1 Evaluation Metric

The evaluation of the models fine-tuned with different datasets has been done using Word Error Rate (WER), which is a standard and widely accepted objective metric in speech recognition (Bălan, 2023).

The formula is defined as:

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

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. The lower the WER is, the better the performance of the model. WER is expressed as a percentage and can exceed 100%, though it has a lower boundary of 0%.

For comparing the different WER between each other, I used a *Relative* WER. This metric is commonly utilized for comparing various models or experiments, as it takes into account the relative nature of the results in relation to a reference point.

$$RelativeWER = \frac{Reference\ WER - Actual\ WER}{Reference\ WER}$$

where *Reference* WER is the WER score that is compared and *Actual* WER corresponds to the WER score measured in a different experiment than the reference.

3.3.2 Statistical Analysis

To evaluate model performance and validate the significance of observed differences, statistical analysis was conducted on the Word Error Rate (WER) across all test samples in the three experimental conditions: no fine-tuning (baseline), fine-tuning on Dutch healthy speech (monolingual), and fine-tuning on English dysarthric speech (cross-lingual).

The Shapiro-Wilk test was applied to assess the normality of WER distributions, and results indicated that the data were not normally distributed. As a result, the Kruskal-Wallis test was used to compare WER values across the three conditions. For pairwise comparisons, the Wilcoxon Signed-Rank test was applied. Additionally, the relative WER difference was used to quantify the performance improvement of each fine-tuned model relative to the baseline. This statistical approach ensures that reported improvements are both robust and meaningful. These tests are based on the research by Su (2024).

The use of statistical tests next to the evaluation metrics ensures that the differences in the performance between the baseline, monolingual and cross-lingual are statistically validated. This approach increases the credibility of the findings and supports the conclusions drawn from the experiments.

3.4 Ethics and Research Integrity

For this thesis, I wanted to make ASR more inclusive for dysarthric users to be used in their daily life. Therefore, I will reflect on some ethical concerns, FAIR principles, Open Science Practices, Bias and Fairness and Reproducibility and Replicability.

3.4.1 Data Ethics and Privacy

Not all the datasets used are open source. As I previously mentioned in Section 3.1.3, for the use of the COPAS corpus it is necessary to request a license via e-mail. In this license it is necessary to explain the reason for the use of the data. All the participants are aware of being recorded and gave permission that their data could be used in research (Van Nuffelen et al., 2009). The rest of the dataset are publicly available, which also received permission for the use of their data (Ons et al., 2014; Rudzicz et al., 2012). For COPAS, the participants are named after their condition together with a number, which protects their privacy. For Domotica, the files of the participants are named after their participant number. The files in the TORGO dataset are separated by sex, together with their participant number. The Common Voice dataset is licensed under CC0, allowing free distribution and adaptation without requiring credit.

3.4.2 FAIR Principles Implementation

Findable. All datasets are easy to find because they are properly referenced in footnotes and described with clear information.

Accessible. The codes are available through the repositories mentioned in this thesis, with access rules in place, and plans for long-term availability. However, the datasets are not available on the GitHub page, since it is not allowed to distribute it myself. For COPAS, creating an account is necessary, as well as signing the terms and conditions via mail.

Interoperable. The dataset is stored in widely-used and accessible formats, including .wav files for audio recordings and .txt files for transcriptions. These formats are supported by most speech processing and machine learning toolkits, making the data easy to integrate into other workflows. Standard metadata terms, such as transcription, are used consistently to describe the data. As I will explain in Section 4, this use of common formats and conventions enhances compatibility with other datasets and tools, supporting future reuse, combination, or benchmarking across research projects.

Reusable. There is clear documentation and information about where the data comes from, so others can use it in the future. The forms for asking the license for using COPAS is on their website.

3.4.3 Bias and Fairness

To address potential biases in this study, several factors are considered. The use of English dysarthric speech for cross-lingual fine-tuning may introduce dataset bias due to linguistic and phonetic mismatches between English and Dutch. Additionally, the small amount of Dutch dysarthric training data may result in unbalanced model performance and poor generalization. Algorithmic fairness is a concern, as the model may perform better on certain speaker characteristics, such as the severity of dysarthria.

3.4.4 Reproducibility and Replicability

To support reproducibility, all code is thoroughly documented and shared on a publicly available GitHub page[11]. There is a step-by-step guide for reproducing the experiments. Some variation in performance is expected due to hardware and training randomness, yet the general findings and experimental procedure are reproducible. Also, no subjective evaluation methods involving human participants were used, which helps avoid ethical concerns and enhances the replicability of the results.

In this section, I described the methodology in this thesis to improve dysarthric speech recognition, as well as the XLSR-53 model in more detail and discussed the evaluation and statistical methods. Through these measures, I ensure that this research takes into account the highest ethical standards, being transparent and making sure that this research is reproducible and replicable, as well as taking into account the potential bias.

[11] https://github.com/AmberL2002/DysarthricASR-cross-lingual-fine-tuning.git

4 Experimental Setup

In order to make this research fully replicable, the experimental setup has to be explained in full detail. The first experiment is running the baseline of XLSR-53 to get the WER with minimal fine-tuning. The second experiment is to discover the performance of healthy mono-lingual fine-tuning. The third experiment is the final experiment, which tests the use of cross-lingual fine-tuning with dysarthric speech. In Section 4.1 I will discuss the preparation of the recordings and the amount of fine-tuning data and testing data. In Section 4.2 I will discuss the design of the experiments. In Section 4.3 I will elaborate on the hyperparameter settings for optimal performance.

4.1 Data Preprocessing

For the TORGO dataset, the recordings were made using two different microphones, which were an array microphone and a head-worn microphone. This captured the same utterances. In this research, the array microphone recordings were used because of their higher clarity and lower electrical noise after sampling (Rudzicz et al., 2012). I excluded recordings, such as '*[say Ah-P-Eee repeatedly]*'. I excluded all files with brackets in the transcripts. In this dataset there were no recordings longer than 10 seconds duration.

For the COPAS dataset, recordings were made using two microphones: a headset and one placed on the table. In the dataset folder, I found only a single set of recordings, with no clear indication of which microphone was used. Therefore, I assume that the two recordings were either merged or only one was retained. I did not use the full dysarthric portion of the dataset, as it contains a large amount of clinical testing data and spontaneous speech. These parts were excluded. Instead, I focused on two repeated sentences that were consistently present across recordings. The dataset is available under a BSD 2-Clause License and the audio files are stored in 16kHz, 16-bit WAV format.

Since the COPAS dataset only includes two repeated sentences per speaker, it was not sufficient on its own for training or evaluating a model on Dutch dysarthric speech. To increase both the amount and variety of speech material, I also included recordings from the Domotica dataset, specifically versions 2, 3, and 4. These sets feature the same 17 dysarthric speakers across different sessions, providing more diverse and representative speech data. I used clean data, which contains recordings made with a close-talk microphone and with background noise removed. Like COPAS, the Domotica data is stored in 16kHz, 16-bit WAV format.

The Common Voice 13.0 dataset, which consists of 3 hours of speech by 53 speakers. For this dataset, there is no standardized microphone, since the voice recordings are contributed by volunteers using their own devices. This was the only dataset in MP3 format, which needed to be converted to wav file. This has been done by using a Python script, as well for removing all capital letters and punctuation marks. This step was especially crucial for normalizing the data, since dysarthric speech frequently contains irregular pauses and stuttering that can result in inconsistent punctuation. Removing punctuation and converting text to lowercase helped for creating a more uniform dataset, enabling the ASR model to concentrate on the essential phonetic content without being affected by transcription variability.

4.1.1 Data Splitting

The datasets needed to be split in separate sets. For the first experiment I only needed the test data since I wanted to get the baseline results of dysarthric speech recognition without fine-tuning. Therefore, I used 30 minutes of testing data. For the mono-lingual experiment, I used 140 minutes of healthy Dutch data to train XLSR-53. Here, I used the same testing set. At last, the cross-lingual experiment, I used the same amount as for the mono-lingual experiment, 140 minutes of dysarthric English data. The reason for using 140 minutes is because the approximately amount of English dysarthric speech dataset TORGO is 200 minutes. As I mentioned in Section 4.1, excluding several recordings causes a certain reduction in the amount of minutes available. Since Shor et al. (2019) also fine-tuned with minimal data, I assumed that multiple hours were not required to achieve meaningful performance improvements.

4.2 Experiment Design

In the first experiment, the goal was to evaluate the baseline performance of the XLSR-53 model on Dutch dysarthric speech without any fine-tuning. For this, only the Dutch dysarthric test set were used. No model adaptation was performed in this condition, allowing for a direct assessment of how well the base model performs on dysarthric Dutch speech.

In the second experiment, the monolingual fine-tuning approach was applied. The XLSR-53 model was fine-tuned using healthy Dutch speech. The goal of monolingual fine-tuning is to see if adapting the XLSR-53 model with healthy Dutch speech improves ASR for Dutch dysarthric speakers by aligning the model to the target language. However, since healthy speech lacks the acoustic patterns of dysarthria, cross-lingual fine-tuning with English dysarthric data may be more effective, as it exposes the model to disordered speech characteristics, helping it better handle similar challenges in Dutch dysarthric speech.

Therefore the third experiment is introduced. Here, English dysarthric speech was used to fine-tune the XLSR-53 model. The same Dutch dysarthric test set was used for evaluation. The results of these experiments are presented in Section 5 where I compared the performances.

I also highlight another experiment in Section 5, I added another experiment which used the fine-tuning data of both monolingual and cross-lingual together to fine-tune the model on combined Dutch typical and English dysarthric data.

4.3 Hyperparameters Setting

The optimal hyperparameters are based on the parameters in Su (2024) and Leivaditi (2023) for a stable model convergence and performance. The following settings were found:

- Max tokens: 3200000
- Optimizer: adam (adam_betas: (0.9, 0.98), adam_eps: 1e08)
- Learning rate (lr): 0.0003
- Scheduler: Fairseq tri_stage scheduler

- Max updates: 10000 steps
- Validation interval: 200
- Gradient accumulation steps: 2

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

5 Results

In this section, I present the results of the three experiments conducted to address the research questions introduced in Section 1.1. Each experiment was designed to evaluate the performance of the XLSR-53 model under different training conditions, with a focus on dysarthric Dutch and English speech. The evaluation metric used throughout is Word Error Rate (WER), which reflects how accurately the model transcribes speech. Statistical tests are used to evaluate the model's performance and validate the significance of observed differences. To decide the correct statistical tests, I used this[12] website to decide the statistical tests. To assess the significance and distribution of the results, I applied the Shapiro-Wilk test to examine normality and the Kruskal-Wallis test for comparing group differences in cases where data did not follow a normal distribution. To identify which groups are significant, I used a Dunn's Post Hoc Test. Together, these experiments aim to provide insight into the impact of fine-tuning strategies on ASR performance for dysarthric speakers. In Section 5.1, I present the performance of all experiments together. In Section 5.2, I present the results of the statistical tests I used to check normality and calculate the significance. The log results are visible in the Appendix section A.

5.1 Performance of Experiments

Experiment 1 aimed to evaluate the high-resource baseline performance of XLSR-53 with testing with dysarthric Dutch speech. In experiment 2, I fine-tuned XLSR-53 with 140 minutes of unseen healthy Dutch speech data. In experiment 3, I fine-tuned XLSR-53 with 140 minutes of dysarthric English speech data. Since the results are not as I expected, I tried to fine-tune the model with the combined data, healthy Dutch and dysarthric English, which is experiment 4. The data for experiment 4 is the same as for the previous experiments, just as the testing data. The resulting Word Error Rates are presented in Table 4 and the relative WERs in Table 5. The results of the relative WER with cross-lingual as reference are negative since there is no improvement. The combined fine-tuned experiment achieved a positive WER for the monolingual and cross-lingual experiments, nonetheless not for the high-resource baseline in Table 6 where the combined experiment is the reference.

Experiment	**WER**
High-resource baseline	82.73
Monolingual	101.44
Cross-lingual	109.51
Combined	93.73

Table 4: Results Third Search Entry

5.2 Statistical results

In Table 7, a Shapiro-Wilk test was conducted to test for normal distribution to decide which parametric or non-parametric test was necessary for calculating significance.

[12]https://www.scribbr.com/statistics/statistical-tests/

Experiment	Relative WER
Baseline	-24.45
Monolingual	-7.37
Combined	-14.42

Table 5: Relative WER with Cross-Lingual as Reference

Experiment	Relative WER
Baseline	-11.74
Monolingual	8.21
Cross-lingual	16.81

Table 6: Relative WER with Combined as Reference

Experiment	Statistic	p-value
Baseline	0.9760	6.39e-05
Monolingual	0.9599	2.36e-07
Cross-lingual	0.5695	1.39e-26
Combined	0.9281	7.41e-11

Table 7: Shapiro-Wilk Test Results

Since all the distributions were not normally distributed, $p < 0.05$, the use of a non parametric Kruskal-Wallis test conducted. The results of this test are shown in Table 8. The Kruskal-Wallis test does show a significant difference for at least one pair of conditions. This test does not show which pairs are significantly different.

Metric	Statistic	p-value
WER	122.5939	2.13e-26

Table 8: Kruskal-Wallis Results

In order 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 are shown in Table 9.

	Baseline	Monolingual	Cross-lingual	Combined
Baseline	1.000000e+00	1.255737e-06	3.285586e-26	8.758655e-03
Monolingual	1.255737e-06	1.000000e+00	1.560395e-07	2.676505e-01
Cross-lingual	3.285586e-26	1.560395e-07	1.000000e+00	2.159391e-13
Combined	8.758655e-03	2.676505e-01	2.159391e-13	1.000000e+00

Table 9: Dunn's Post Hoc Test

A significance level of 0.05 is used for the Dunn's post hoc test[13] and the results are that the baseline is significantly different ($p = 1.3 \times 10^{-6}$), ($p = 3.3 \times 10^{-26}$) than monolingual and cross-lingual experiments, which indicates that the baseline outperforms the other experiments. For the monolingual and combined experiments, the WERs are not significantly different $p = 0.268$ as well for the baseline and combined experiments ($p = 0.009$).

Four experiments evaluated XLSR-53's performance on dysarthric Dutch speech (Table 4 and Table 5. The high-resource baseline (Experiment 1) performed best with a WER of 82.73, while monolingual and cross-lingual fine-tuning (Experiments 2 and 3) yielded higher WERs of 101.44 and 109.51. The combined approach (Experiment 4) showed partial improvement (WER = 93.73) but did not outperform the baseline. Relative WER in Table 5) confirmed the cross-lingual model underperformed across comparisons. As the Shapiro-Wilk test indicated non-normal distributions (Table 7), a Kruskal-Wallis test was used and revealed significant differences $p = 2.13 \times 10^{26}$ in Table 8. Dunn's post hoc test (Table 9 showed that the baseline differed significantly from all other models, except combined $p = 0.009$. Cross-lingual results also differed significantly from monolingual and combined models. Only the monolingual and combined models did not show a significant difference $p = 0.268$.

[13] https://www.adventuresinmachinelearning.com/mastering-kruskal-wallis-and-dunns-test-a-comprehensive-guide/

6 Discussion

This section provides an interpretation of the experimental outcomes and evaluates the performance of the XLSR-53 model under the different fine-tuning strategies. Specifically, it compares the impact of mono-lingual fine-tuning on healthy Dutch speech versus cross-lingual fine-tuning using English dysarthric data. The results are discussed in relation to the research question and hypotheses presented earlier. Additionally, potential explanations for the observed effects, the influence of dataset characteristics, and broader implications for ASR in dysarthric speech are addressed.

The main goal of this study was to investigate whether cross-lingual fine-tuning with English dysarthric speech improves recognition performance for Dutch dysarthric speech, compared to using mono-lingual fine-tuning on healthy Dutch speech. This leads to the central research question:

> **Does cross-lingual fine-tuning with English dysarthric speech, instead of monolingual fine-tuning with healthy speech, improve the performance of the self-supervised model XLSR-53 for Dutch dysarthric speech in ASR?**

The hypothesis was that multilingual exposure allows the model to generalize across phonetic variability, making it more robust to atypical pronunciations common in dysarthria, a claim supported in earlier work such as Hernandez et al. (2022) and P. Wang and Van Hamme (2023).

6.1 Performance of High Resource Baseline

The surprisingly good performance of the high-resource baseline XLSR-53 model on Dutch dysarthric speech suggests that its multilingual pre-training already provided it with a useful foundation for handling non-standard pronunciation. One likely reason is that the model had exposure to Dutch and Dutch-like phonetic patterns through its training on large, diverse datasets such as Common Voice 6.1, which includes contributions from a range of Dutch dialects (Hernandez et al., 2022). The amount of 32 hours of fine-tuning data may have helped the model generalize more effectively to the dysarthric Dutch speech used in this study. In contrast, the models fine-tuned on just 140 minutes of data in this thesis, whether healthy Dutch or dysarthric English, were working with much more limited information. According to Javanmardi et al. (2024), models that begin from robust, extensively pre-trained architectures generally achieve better performance, especially in scenarios where fine-tuning data is limited. This is supported by findings from Shor et al. (2019), who demonstrated that pre-trained models can still perform well on dysarthric speech with very limited fine-tuning, highlighting the strength of self-supervised learning in low-resource conditions. However, the model by Shor et al. (2019) was not XLSR-53. In theory, cross-lingual fine-tuning would have a positive effect.

6.2 Cross-lingual Fine-Tuning vs. Mono-Lingual Fine-Tuning

Another explanation lies not in utterance length, but in the phonetic confusability of dysarthric speech. Although the English TORGO dataset contains mostly isolated words and short phrases, many of the longer Dutch test and training utterances in this study were still frequently not correctly recognized. This aligns with findings from the STARDUST project (**?**), which showed that severe dysarthria often reduces the number of distinguishable phonetic tokens a speaker can produce. As

a result, short and long utterances alike can be problematic when phonetic contrasts are weak or inconsistent. Similarly, Mengistu and Rudzicz (2011) reported that 83% of recognition errors occurred in single-word utterances, largely due to the presence of homophones created by articulation errors such as deletions and substitutions. However, they also showed that pronunciation lexicon adaptation (PLA) was significantly more effective in longer utterances, as context helps resolve ambiguity. In this thesis, it is likely that phoneme-level confusion and pronunciation variability, rather than utterance length alone, contributed to the high WER observed in both short and long sentences.

The combined fine-tuning approach, which included both Dutch and English data, outperformed the purely cross-lingual model, likely because it offered a more balanced exposure to both target language features and general phonetic variability. This aligns with findings from Hernandez et al. (2022), who demonstrated that multilingual fine-tuning using XLSR-53 improves dysarthric speech recognition across multiple languages, including English, Spanish, and Italian. Their results showed that XLSR-based features consistently outperformed monolingual models, suggesting that cross-lingual representations are more robust to variation in impaired speech. Also, Shor et al. (2019) also found that combining dysarthric and accented speech data led to further improvements, especially in diverse or noisy acoustic conditions.

In this thesis, the combined model may have benefited from a similar effect: English dysarthric data introduced articulation variability, while the Dutch data, despite dialectal variation, helped ground the model in the target language. Although the combined approach did not surpass the high-resource baseline, it clearly performed better than cross-lingual fine-tuning alone, supporting the conclusion by Hernandez et al. (2022) that strategically combining multilingual and in-language data enhances robustness for dysarthric ASR systems.

6.3 Additional Observations

Age Differences Between Datasets. The English and Dutch datasets differed considerably in the age of speakers, which could affect the acoustic characteristics of speech. Su (2024) highlighted the importance of using age-matched training data for improved model performance on impaired speech. I did not include age-matched datasets in the current research, potentially leading to mismatches in vocal characteristics. As I mentioned in Section 2.2.1 and Section 3.1.3, the age ranges were for Dutch training data unknown, for English training data between 16 and 50 years old and for testing data between 14 and 80 years old (with COPAS and Domotica combined).

Determining Which Language. Although cross-lingual models are often praised for their ability to generalize across languages, their effectiveness strongly depend on the degree of similarity between the source and target domains. This includes not just the language itself, but also the speech style, prosody, phonotactics, and even demographic factors such as speaker age or regional accent. As demonstrated by Javanmardi et al. (2024), fine-tuning with a small amount of dysarthric data can yield meaningful improvements only when there is sufficient alignment between the characteristics of the fine-tuning data. Using English dysarthric speech to improve Dutch dysarthric recognition could be a weak alignment between speech characteristics and creates differences in phoneme distributions, lexical stress patterns, or sentence structure may limit the transferability of learned representations. This issue becomes even more pronounced when the fine-tuning dataset is small, as the model lacks the exposure needed to adjust effectively to mismatched features. Moreover,

impaired speech introduces another layer of variability, often with speaker-specific articulatory patterns, which may not be captured adequately through generalized multilingual fine-tuning. Therefore, while cross-lingual models show promise for low-resource ASR tasks, their success largely depends on how closely the source language and speech style match the target task.

Speaker-dependent models. Because of the variation between speakers in both the fine-tuning and testing datasets, such as differences in age, demographics, sentence complexity, and severity levels of dysarthria, the model may struggle to generalize. As Rosen and Yampolsky (2000) explained, speaker-dependent ASR systems typically achieve higher accuracy when matched to an individual user's speech patterns. These systems use the speaker's own templates, making them more tolerant of atypical but consistent speech. However, when fine-tuned on one set of speaker characteristics and tested on another, performance can decline, especially if the speech patterns differ significantly. This mismatch likely contributed to the reduced performance in this study.

6.4 Limitations

This study faced several limitations that are important to acknowledge when replicating this study.

A key limitation was the restricted availability and diversity of dysarthric speech datasets. Public resources such as TORGO, COPAS, and Domotica contain relatively few speakers, and much of the speech consists of structured or repeated utterances. This limited variety makes it difficult for the model to generalize to broader, more realistic scenarios, especially for individuals with rare articulation patterns or highly variable symptoms. The small amount of data also increases the risk of overfitting during fine-tuning, particularly when trying to adapt to the unique characteristics of dysarthric speech.

Another limitation concerns cross-lingual phonetic mismatch. Although XLSR-53 is pre-trained on a wide range of languages, fine-tuning it on English dysarthric speech and evaluating it on Dutch dysarthric speech introduces inconsistencies at the phonetic level. Differences in phoneme inventories, intonation patterns, and typical pronunciation errors between English and Dutch could prevent the model from effectively learning the features most relevant to Dutch dysarthric speech. As this thesis has shown, these differences may have limited the potential benefit of cross-lingual fine-tuning.

The study also relied solely on Word Error Rate (WER) as the evaluation metric. WER calculates the proportion of word-level errors by counting the number of substitutions, insertions, and deletions needed to transform the predicted transcript into the reference transcript, divided by the number of words in the reference. While this is a standard metric in ASR, it does not always reflect how understandable or useful the output is to a listener. For example, a predicted sentence that is phonetically close to the reference but structurally different might still result in a WER of 100%, even though it is more intelligible than another output with a lower WER but more distortion. This occurred multiple times in the cross-lingual model, where minor pronunciation errors led to worse scores than completely unintelligible outputs. WER does not account for phoneme-level closeness or partial recognition, which limits its ability to capture subtle improvements in intelligibility that could be meaningful in real-world applications.

Additionally, the study did not incorporate age-matched fine-tuning data. Su (2024) has shown

that age-related differences in vocal characteristics, such as pitch, speaking rate, and articulation can affect ASR performance. By fine-tuning only on healthy adult speech with unspecified age ranges, the model may have missed important acoustic patterns that align more closely with the target dysarthric speech, which often comes from older speakers.

As I mentioned before, according to P. Wang and Van Hamme (2023) the use of dialectal variations in fine-tuning can reduce generalizability. This can be solved according to Shor et al. (2019) by focusing on layers closest to the input, like the encoder. In my current setup, I did not specifically focus on the layers closest to the input. I replaced and fine-tuned the final classification head and froze only the feature extractor, which includes the initial convolutional layers. However, the rest of the encoder layers remained trainable, meaning the model updated weights across all higher layers during fine-tuning. According to Shor et al. (2019), focusing on the lower encoder layers could improve generalization in cross-lingual settings, especially when dealing with dialectal variation. To potentially lower the WER, I recommend an experiment with freezing the higher layers and fine-tuning only the lower ones, or apply layer-wise learning rates that emphasize updates closer to the input.

Finally, the study did not explicitly take into account dysarthria severity. Each dataset used in the experiments includes a mixture of mild, moderate, and severe cases of dysarthria. However, the data was not separated based on severity levels, so it is unclear how much of the training and testing material came from speakers with mild, moderate, or severe dysarthria. A potential imbalance is critical, as ASR systems tend to perform substantially worse on speech from individuals with severe dysarthria (Jaddoh et al., 2023; Young & Mihailidis, 2010).

Thus, limitations in dataset size, language matching, evaluation metrics, speaker demographics, and severity distribution all impacted the outcomes of this research. Addressing these limitations could be useful for future research, which I will discuss in Section 7.2.

7 Conclusion

This thesis explored whether cross-lingual fine-tuning on dysarthric speech can help improve recognition. While past work showed benefits from multilingual fine-tuning, it is still unclear if cross-lingual fine-tuning helps with limited dysarthric data. This study aimed to address that gap and improve recognition for Dutch dysarthric speech. In Section 7.1 I will discuss shortly the key findings and how the results contradicted my hypotheses. In Section 7.2, I will discuss potential future research based on my limitations addressed earlier. In the final Section 7.3, I will discuss the potential applications this thesis has.

7.1 Summary of the Main Contributions

This thesis examined whether cross-lingual fine-tuning with English dysarthric speech can improve ASR performance on Dutch dysarthric speech, compared to monolingual fine-tuning using healthy Dutch speech. The self-supervised XLSR-53 model was evaluated under four experimental conditions, including a high-resource baseline, monolingual, cross-lingual, and combined fine-tuning strategies. Key findings include:

- The high resource baseline model, performed best overall. This suggests that the fine-tuning multilingual model already had strong generalization abilities for dysarthric Dutch speech. Both monolingual and cross-lingual fine-tuning led to worse performance, contradicting the initial hypothesis that introducing phonetic variability would help the model adapt to impaired speech.
- Neither monolingual fine-tuning on healthy Dutch speech nor cross-lingual fine-tuning on English dysarthric speech led to improved performance. This outcome contradicts the original hypothesis that cross-lingual fine-tuning would help the model handle atypical pronunciation more effectively.
- The combined fine-tuning approach, which merged Dutch and English data, performed better than cross-lingual fine-tuning alone, but still failed to surpass the high resource baseline. This suggests that combining target-language and variability-rich data can help, but is insufficient without further alignment between training and test conditions.

Overall, this study contributes to the understanding that while cross-lingual strategies hold promise, their success depends heavily on the alignment of linguistic, acoustic, and demographic features.

7.2 Future Work

The limitations I mentioned in Section 6.4, can also build on future research.

Future research should focus on using larger and more diverse dysarthric speech datasets. Current datasets are small and repetitive, which limits model generalisation. Including more speakers with varying types and severity levels of dysarthria could make models more robust and useful.

Further work could also explore how speech sounds differ across languages. Since cross-lingual fine-tuning did not help in this study, focusing on shared phonemes or using phoneme-based metrics

like Phoneme Error Rate (PER) might improve performance. PER is especially useful for identifying subtle pronunciation errors common in dysarthric speech.

In addition to WER, future evaluations should consider other metrics like intelligibility or task success to better capture real-world effectiveness.

It may also help to separate data by severity and train different models for mild, moderate, and severe dysarthria. Finally, experiments with freezing higher layers and fine-tuning lower ones could improve cross-lingual generalisation, especially with limited data.

7.3 Impact and Relevance

Although this study did not lead to improved recognition performance for dysarthric Dutch speech, it offers valuable insights into why current ASR approaches may fall short and what factors need to be addressed in future development. By systematically testing monolingual, cross-lingual, and combined fine-tuning strategies, this research highlights key challenges in adapting existing models to non-standard speech, such as data mismatch, lack of speaker diversity, and the limitations of commonly used evaluation metrics like Word Error Rate.

These insights are particularly relevant for real-world applications like speech-based Personal Emergency Response Systems (PERS). Traditional PERS often rely on physical buttons, which can be difficult or inaccessible for people with limited mobility. A voice-controlled alternative could allow users to call for help more easily, reduce stigma, and even lower healthcare costs. However, such systems must be able to understand speech that may be impaired, aged, or spoken under stress—conditions that standard ASR systems often fail to handle reliably.

By identifying the gaps and limitations in current ASR strategies for dysarthric speech, this thesis helps pave the way for more inclusive and accessible speech technology. It helps clarify what challenges need to be solved to make speech technology more reliable for people with speech impairments and highlights the need for future research that focuses on specific speaker needs and smarter use of limited data.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... others (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Bălan, D. A. (2023). Improving the state-of-the-art frisian asr by fine-tuning large-scale cross-lingual pre-trained models.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... others (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, *12*(2), 246–269.

Doyle, P. C., Leeper, H. A., Kotler, A.-L., Thomas-Stonell, N., O'Neill, C., Dylke, M.-C., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, *34*, 309–316.

Espana-Bonet, C., & Fonollosa, J. A. (2016). Automatic speech recognition with deep neural networks for impaired speech. In *Advances in speech and language technologies for iberian languages: Third international conference, iberspeech 2016, lisbon, portugal, november 23-25, 2016, proceedings 3* (pp. 97–107).

Fan, R., Shankar, N. B., & Alwan, A. (2024). Benchmarking children's asr with supervised and self-supervised speech foundation models. *arXiv preprint arXiv:2406.10507*.

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, *11*(3), 165–175.

Gales, M. J., Knill, K. M., Ragni, A., & Rath, S. P. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. , 16–23.

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

Grosman, J. (2021). *Fine-tuned XLSR-53 large model for speech recognition in Dutch.* `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-dutch`.

Hernandez, A., Pérez-Toro, P. A., Nöth, E., Orozco-Arroyave, J. R., Maier, A., & Yang, S. H. (2022). Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv preprint arXiv:2204.01670*.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451–3460.

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Commu-*

nication, *16*(3), 186–196.

Jaddoh, A., Loizides, F., & Rana, O. (2023). Interaction between people with dysarthria and speech recognition systems: A review. *Assistive Technology*, *35*(4), 330–338.

Javanmardi, F., Kadiri, S. R., & Alku, P. (2024). Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech. *IEEE journal of biomedical and health informatics*.

Joy, N. M., & Umesh, S. (2018). Improving acoustic models in torgo dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(3), 637–645.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J. R., Huang, T. S., Watkin, K. L., ... others (2008). Dysarthric speech database for universal access research. , *2008*, 1741–1744.

Kim, M., Kim, Y., Yoo, J., Wang, J., & Kim, H. (2017). Regularized speaker adaptation of kl-hmm for dysarthric speech recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(9), 1581-1591. doi: 10.1109/TNSRE.2017.2681691

Leivaditi, S. (2023). The role of speech elicitation methods and disease factors in dysartrhric asr system development.

Mann, W. C., Belchior, P., Tomita, M. R., & Kemp, B. J. (2005). Use of personal emergency response systems by older individuals with disabilities. *Assistive technology*, *17*(1), 82–88.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). The nemours database of dysarthric speech. , *3*, 1962–1965.

Mengistu, K. T., & Rudzicz, F. (2011). Adapting acoustic and lexical models to dysarthric speech. In *2011 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4924–4927).

Ons, B., Gemmeke, J. F., & hamme, H. V. (2014). The self-taught vocal interface. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*, 1–16.

Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology: The stardust project. *Clinical linguistics & phonetics*, *20*(2-3), 149–156.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Qian, Z., & Xiao, K. (2023). A survey of automatic speech recognition for dysarthric speech. *Electronics*, *12*(20), 4278.

Rietveld, T., & Van Heuven, V. J. (2016). *Algemene fonetiek (4e geheel herziene druk)*. Bussum: Coutinho.

Rosen, K., & Yampolsky, S. (2000). Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, *16*(1), 48–60.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, *46*, 523–541.

Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., ... others (2019). Personalizing asr for dysarthric and accented speech with limited data. *arXiv preprint arXiv:1907.13511*.

Su, C. (2024). Enhancing english dysarthric speech recognition with age-matched healthy speech: A fine-tuning approach using wav2vec 2.0.

Van Nuffelen, G., De Bodt, M., Middag, C., & Martens, J.-P. (2009). Dutch corpus of pathological and normal speech (copas). *Antwerp University Hospital and Ghent University, Tech. Rep*.

Vaquero, C., Saz, O., Lleida, E., Marcos, J., Canalís, C., & De Educación, C. P. (2006). Vocaliza: An application for computer-aided speech therapy in spanish language. *IV Jornadas en Tecnología*

del Habla, 321–326.

Wang, H., Jin, Z., Geng, M., Hu, S., Li, G., Wang, T., ... Liu, X. (2024). Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 12311–12315).

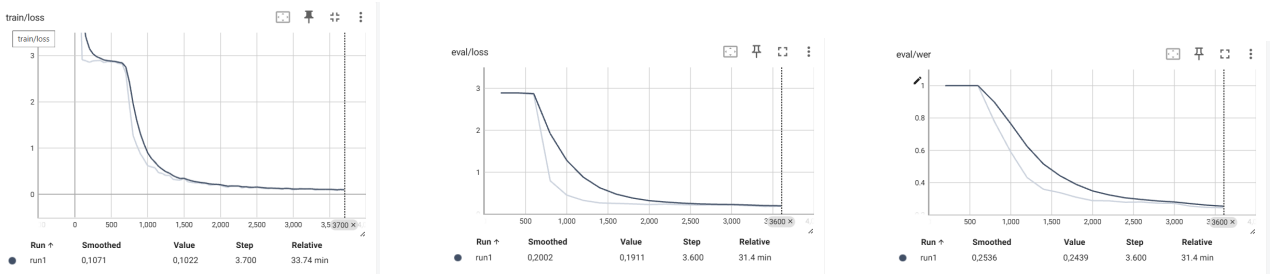
Wang, P., & Van Hamme, H. (2023). Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of dutch dysarthric speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*(1), 15.

World Health Organization. (2024, March 14). *Over 1 in 3 people affected by neurological conditions, the leading cause of illness and disability worldwide.* Retrieved from https://www.who.int/news/item/14-03-2024-over-1-in-3-people-affected-by-neurological-conditions--the-leading-cause-of-illness-and-disability-worldwide (Accessed: 2025-03-12)

Young, V., & Mihailidis, A. (2010). 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.

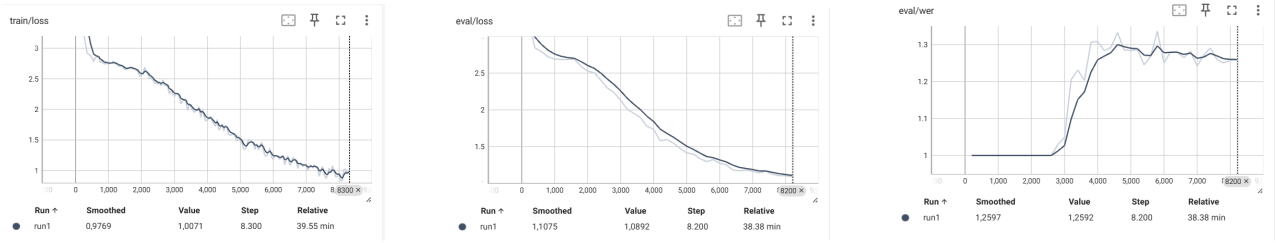
Appendices

A Loss and WER dynamics



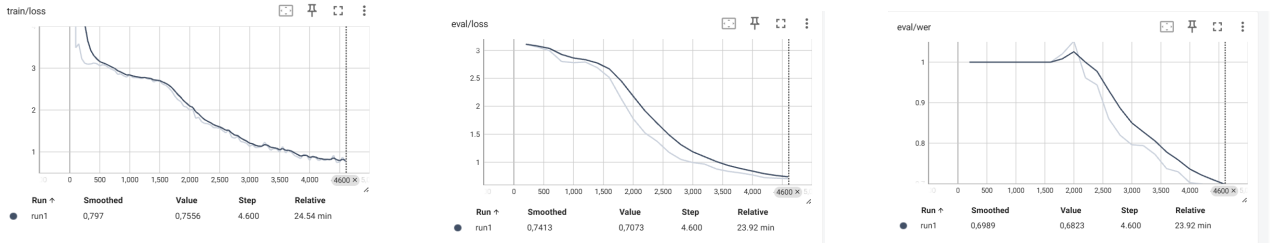
(a) Train Loss (b) Eval Loss (c) Eval WER

Figure 2: Fine-tuned with Dutch typical speech



(a) Train Loss (b) Eval Loss (c) Eval WER

Figure 3: Fine-tuned with English dysarthric speech



(a) Train Loss (b) Eval Loss (c) Eval WER

Figure 4: Fine-tuned with combined speech

B AI tools in Master Thesis

Acknowledging, citing and referencing use of AI tools and technologies in the Master thesis

For MSc Voice Technology students:

AI tools may be used for the following support functions without disclosure: grammar checking, spell checking, translation between languages, generating practice questions for self-assessment, and clarifying publicly available technical concepts. These functions must not alter the substance, structure, or argumentation of your work.

Use of AI for generating code, algorithm explanations, experimental design, or data interpretation must be disclosed. Students must demonstrate understanding of AI model limitations relevant to their research domain and justify their choice of AI tools over alternatives.

Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used ChatGPT for the following purpose: giving stylistic feedback and checking grammar for the introduction, literature review, discussion and conclusion. Also my title is generated by AI. My original title was: Cross-Lingual Fine-Tuning for Improving Dysarthric Speech Recognition, which was the prompt for creating the title. All content was subsequently reviewed, verified, and substantially modified by me.

Amber Lankheet / 10-06-2025

Prohibited uses include: AI generation of research hypotheses, experimental methodology, data analysis interpretations, conclusions, or any content where independent reasoning and domain expertise are being assessed. Submitting AI-generated text without disclosure constitutes academic misconduct