# Improving State-of-the-Art ASR Systems for Speakers with Dysarthria

## Applying Low-Rank Adaptation Transfer Learning to Whisper

**Mirella Günther**[1]
**Supervisors: Zhengjun Yue**[1]**, YuanYuan Zhang**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 20, 2024

Name of the student: Mirella Günther
Final project course: CSE3000 Research Project
Thesis committee: Zhengjun Yue, YuanYuan Zhang, Thomas Durieux

# Abstract

Dysarthria is a speech disorder that limits an individual's ability to clearly articulate, due to the weakening of the muscles involved in speech. Despite recent advances in Automatic Speech Recognition (ASR), the recognition of dysarthric speech remains a significant challenge because of the limited availability of dysarthric speech data, significant speaker variability, and the mismatch between typical and dysarthric speech patterns. This paper addresses these challenges by using transfer learning and Low-Rank Adaptation (LoRA) techniques to enhance the performance of the state-of-the-art ASR model Whisper on dysarthric speech. By fine-tuning Whisper with the TORGO dataset, this study aims to adapt the pre-trained models to better recognise dysarthric speech patterns, thus reducing Word Error Rates (WER) and improving accessibility for individuals with speech impairments. Experimental results indicate that this approach can improve speech recognition performance since the Large-V2, Large-V3 and the corresponding distilled models achieved a reduction in WER after fine-tuning. The Large-V3 model achieved the greatest relative WER reduction of 22.65%.

**Index Terms**: Automatic Speech Recognition, Dysarthria, Transfer Learning, Low-Rank Adaptation, Whisper Model

# 1. Introduction

Dysarthria is a neuromotor speech disorder that often arises in conditions such as Cerebral Palsy and Amyotrophic Lateral Sclerosis (ALS). Characterised by the weakening, slowing, or lack of coordination of the muscles involved in speech production, dysarthric speech typically manifests through abrupt shifts in pitch, diminished articulation, and a breathy quality [1, 2]. Individuals with dysarthric speech frequently experience additional physical impairments that complicate the use of traditional technological interfaces like mouse and keyboard. Consequently, voice technology has the potential to enhance the independence and quality of life for individuals with physical disabilities [3].

Voice technology relies on Automatic Speech Recognition (ASR), defined as "the process and the related technology for converting [a] speech signal into its corresponding sequence of words or other linguistic entities by means of algorithms" [4, p. 1]. While ASR technologies hold significant promise for individuals with dysarthric speech, research indicates that atypical speech recognition encounters three primary challenges within ASR systems [5]. First, there is generally less atypical speech data available, leading to data scarcity. This limits the ability to train robust ASR models specifically tailored for atypical speech recognition. Second, there is considerable speaker variability within atypical speech, including varying severity levels of dysarthria, complicating the generalisation of ASR models across different speakers [6]. Third, there is a mismatch between typical and atypical speech, making it difficult for models trained on typical speech data to accurately recognise dysarthric speech.

Due to these challenges, ASR systems tend to achieve lower accuracy when used by individuals with speech pathologies [7]. Therefore, it is crucial to fine-tune these technologies to accommodate dysarthric speech, thereby enabling individuals with dysarthria to benefit fully from ASR. Several solutions have been proposed to address these challenges. For instance, data augmentation techniques have been explored to simulate dysarthric speech using other, more widely available speech, mitigating the data scarcity problem [8, 9]. Moreover, the Voiceitt app [10] allows dysarthric speakers to train a model with their unique word pronunciations by reading out a series of phrases, addressing both speaker variability and the mismatch between typical and dysarthric speech.

Another promising approach to overcoming data scarcity and speech mismatches is transfer learning, which simplifies learning for a new task based on knowledge from a similar task [11]. Previous studies have used transfer learning to improve recognition of dysarthric speech. For example, [12] proposed a transfer learning method for recognising Japanese dysarthric speech by pre-training on Japanese non-dysarthric speech and non-Japanese dysarthric speech, and then fine-tuning on the target Japanese dysarthric speech. This approach achieved a 33.3% relative improvement in Phoneme Error Rate (PER) compared to random initialisation and a 15.3% relative improvement compared to non-Japanese dysarthric speech being excluded from pre-training. Similarly, another study compared the performance of the state-of-the-art ASR model Whisper [13] with a hybrid ASR model for Dutch dysarthric speech, finding that fine-tuning Whisper with dysarthric data led to Word Error Rate (WER) improvements of 7.3% for moderate severity and 10.7% for severe severity levels, although it did not outperform the hybrid model overall [14]. Additionally, [15] focused specifically on fine-tuning Whisper's encoder module alongside a classifier to investigate the speech recognition of dysarthric speech. An average word recognition accuracy of 59.78% was achieved on the UA-Speech Corpus [16], demonstrating the potential of Whisper in handling dysarthric speech.

This study aims to evaluate the speech recognition capabilities of Whisper when further trained with dysarthric speech data. Whisper was chosen as the primary model for this investigation because it is known for its robust performance across various speech inputs [13], therefore offering a promising starting point for adapting ASR technology to accommodate dysarthric speech patterns. By fine-tuning the entire model using low-rank adaptation transfer learning techniques, this research seeks to address the research question: **How can Automatic Speech Recognition models designed for typical speech be adapted using fine-tuning to better recognise dysarthric speech?** The Whisper model will be re-trained on a dataset of dysarthric speech to enable individuals with dysarthria to benefit from state-of-the-art ASR technology. Furthermore, the same approach will also be applied to distilled versions of the Whisper model, which are more compact and therefore potentially easier to integrate into software. This research will address the following sub-questions:

- How can pre-trained models be leveraged to improve the recognition of dysarthric speech?
- How well does Whisper, after low-rank adaptation fine-tuning, generalise to different types and severities of dysarthric speech, as well as to typical speech?
- How does the performance of the distil-Whisper models compare to that of the standard Whisper models after applying low-rank adaptation fine-tuning?

In Section 2, the Whisper and distil-Whisper models are described, and the methodology is established. Following this, Section 3 outlines the experimental setup for implementing transfer learning on these models. Section 4 presents the results, which are then further discussed in Section 5. Section 6 summarises the key findings and suggests areas for future research. Finally, Section 7 explores the ethical considerations within this research.

# 2. Methodology

This section provides insight into the methodology that will be used throughout this study. This research will use Whisper, an open-source, state-of-the-art speech recognition model described in Subsection 2.1. Furthermore, the distil-Whisper models will also be used throughout this research, as established in Subsection 2.2. Section 2.3 establishes how transfer learning will be used to fine-tune the aforementioned models, and Subsection 2.4 explains how the performance of the models will be evaluated.

## 2.1. Whisper

Whisper is an ASR model that was published by OpenAI in September 2022 [13]. It has been trained on 680,000 hours of a wide range of labelled data allowing it to generalise well to a wide range of speech in a zero-shot transfer setting. This can be seen in the performance measures presented in [13].

The original Whisper family contains five Whisper models, named after their increasing number of parameters: Tiny, Base, Small, Medium, and Large, which will be referred to as Large-V1. The models differ in terms of their trainable parameter counts and the number of transformer encoder-decoder layers they employ, as outlined in Table 1. There are two versions for each of these models, one trained with only English speech and one trained with multilingual speech. The exception to this is the Large-V1 model, which has no separation of English and other languages.

After the original release of these Whisper models, the Large-V1 model was trained for an additional 2.5 times the original epochs and was subsequently denoted as the Large-V2 model. In November 2023, the Large-V3 model was created, which was trained for 2 epochs on a dataset consisting of 1 million hours of weakly labelled audio and 4 million hours of audio labelled by Whisper Large-V2 [17].

Table 1: *Whisper Model Family Specifications* [13, 18]

| Model | Layers | Width | Heads | Parameters |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39M |
| Base | 6 | 512 | 8 | 74M |
| Small | 12 | 768 | 12 | 244M |
| Medium | 24 | 1024 | 16 | 769M |
| Large-V1 | 32 | 1280 | 20 | 1550M |
| Large-V2 | 32 | 1280 | 20 | 1550M |
| Large-V3 | 32 | 1280 | 20 | 1550M |
| D-S | 12 | 768 | 12 | 166M |
| D-M | 24 | 1024 | 16 | 394M |
| D-L-V2 | 32 | 1280 | 20 | 756M |
| D-L-V3 | 32 | 1280 | 20 | 756M |

## 2.2. Distilled Whisper Models

In machine learning, a distilled model refers to a smaller model that has inherited knowledge from a larger model. [18] was able to distil the Small, Medium, Large-V2 and Large-V3 Whisper models into models with fewer parameters, as shown in Table 1. These distilled models are 5.8 times faster than their corresponding standard Whisper models and achieve a performance of within 1% of Whisper's WER. Such compact models are particularly beneficial for deployment in low-latency or resource-limited settings. Therefore, this research will investigate the application of transfer learning to these smaller models. The aim is to explore whether, despite having fewer parameters, these models can still be fine-tuned to improve their recognition of dysarthric speech.

## 2.3. Transfer Learning

Given the significant mismatch between typical and dysarthric speech, transfer learning will be employed. Transfer learning in this context involves taking the pre-trained Whisper and distil-Whisper models, which have demonstrated proficiency in typical speech recognition, and further training them with dysarthric speech data. This approach enables the models to adapt to the unique characteristics of dysarthric speech, improving their overall performance in recognising such speech patterns.

By using pre-existing knowledge from the typical speech recognition domain, transfer learning facilitates more effective learning in the dysarthric speech domain. Instead of starting from scratch and training a model solely on dysarthric data, which may require a large number of labelled samples to achieve good performance, transfer learning reduces the amount of data required for training and accelerates the learning process. This ultimately improves the model's ability to recognise dysarthric speech patterns despite data scarcity and domain mismatch.

To efficiently handle the extensive parameter space of Whisper models, particularly the Large-V1, Large-V2 and Large-V3 models which contain 1.5 billion parameters each, Low-Rank Adaptation (LoRA) [19] will be used for parameter-efficient fine-tuning. LoRA operates on the principle of low-rank decomposition of weight updates, significantly reducing the parameter space. Instead of updating the entire weight matrix, LoRA updates two smaller matrices that approximate the original updates.

Given a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA represents the update as a product of two low-rank matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where $r$ is much smaller than $d$ and $k$. The update $W_0 + \Delta W$ is thus decomposed into $W_0 + BA$. During training, the original weights $W_0$ are frozen, and only $A$ and $B$ are updated. The training process involves creating the low-rank matrices $A$ and $B$ which are initialised randomly. The model then updates these matrices during fine-tuning. The output of a forward pass is modified to include the low-rank update:

$$h = W_0x + \Delta Wx = W_0x + BAx. \tag{1}$$

This allows the model to adapt to new tasks efficiently, as the number of trainable parameters is drastically reduced compared to full fine-tuning [19]. From now onwards, whenever fine-tuning is mentioned, the assumption can be made that this refers to fine-tuning using LoRA.

The implementation of LoRA will be facilitated through the PEFT (Parameter-Efficient Fine-Tuning) library [20]. This library provides tools for applying LoRA and other parameter-efficient techniques to large-scale models, making it easier to integrate these approaches into existing training pipelines. By employing LoRA, the project aims to achieve significant improvements in dysarthric speech recognition while optimising resource usage.

### 2.4. Evaluation Metric

To evaluate the performance of the Whisper models and their distilled counterparts, Word Error Rate (WER) will be used as the primary metric. WER is based on the Levenshtein distance [21] and is a standard measure in the field of speech recognition that quantifies the accuracy of the transcriptions produced by an ASR system. It provides a clear, numerical representation of how well the model performs in recognising and transcribing spoken language. It can be calculated using the following formula:

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \tag{2}$$

where:

- $S$ is the number of substitutions,
- $D$ is the number of deletions,
- $I$ is the number of insertions,
- $N$ is the number of words in the reference.

This means that a lower WER indicates a closer match to the reference transcript. To quantify the change in WER before and after fine-tuning, Relative Word Error Rate Difference will be used. The equation for this is as follows:

$$\text{Relative WER Difference} = \frac{\text{WER}_{\text{after}} - \text{WER}_{\text{before}}}{\text{WER}_{\text{before}}} \tag{3}$$

A positive value can therefore be interpreted as an increase in WER, whilst a negative value indicates a decrease in WER.

## 3. Experimental Setup

This section provides an overview of the conducted experiments. It starts with a detailed description of the chosen dataset, as outlined in Subsection 3.1. Following this, Subsection 3.2 presents the specific experiments performed.

### 3.1. The TORGO dataset

*3.1.1. Dataset description*

In this work, the TORGO dataset [22] was used, a collaborative effort between the University of Toronto and the Holland-Bloorview Kids Rehab Hospital in Toronto. This English-speech database comprises recordings from eight individuals (five male and three female) with dysarthric speech impairments due to cerebral palsy or ALS. Furthermore, eight age- and gender-matched control participants are also included. The subjects with dysarthria can be categorised into four different severity levels, as summarised in Table 2. Since only one subject is in the 'moderate' category, after reviewing the audio, I have decided to additionally categorise M05 as 'moderate' instead of 'moderate-to-severe'.

Table 2: *Patient Dysarthria Severity [23]*

| Severity | Subjects |
|---|---|
| Very Mild | F04, M03 |
| Moderate | F03 |
| Moderate-to-Severe | M05 |
| Severe | F01, M01, M02, M04 |

The participants were instructed to read text displayed on a screen, which consisted of various stimuli, namely non-words, short words, restricted sentences and unrestricted sentences. Non-words were included to assess the baseline abilities of the speakers and consisted of repetitions of various phonetic sequences. A wide range of short words were included, as they don't require word boundary detection. Additionally, subjects were presented with phoneme-rich restricted sentences and they described various images to produce unrestricted sentences. Each participant has about 3 hours of recorded data, which is comprised of 500 utterances from the dysarthric speakers, and 1200 utterances from the speakers without dysarthria [24].

*3.1.2. Division into Test and Training Set*

The TORGO dataset lacks a predefined division into test and training sets. Most studies on the TORGO dataset employ the leave-one-speaker-out (LOSO) method, where one speaker is isolated as the test set while the others form the training set [23, 25, 26, 27]. Other common divisions include an n-fold cross-training strategy, as detailed in [28], or a random division of all the audio files [29].

However, it is important to note that all participants were given the same prompts, resulting in significant repetition across different individuals. Specifically, when a LOSO out approach is used, depending on which subject makes up the test set, there will be a prompt overlap with the training set of at least 95.7% [30]. With only eight dysarthric speakers and substantial overlap in their utterances, the aforementioned approaches tend to overestimate model performance due to their narrow focus on TORGO's specific phrases. This issue is further amplified by Whisper's strong language model, which can recognise the repetition of prompts, leading to an inflated performance evaluation.

To address these limitations, I decided to split the data based on the prompts given to the subjects, rather than dividing the subjects themselves. First, I removed the typical speech from the dataset to focus solely on training Whisper with dysarthric data, since it has already been trained on typical speech. From this dysarthric dataset, I separated 70% of the prompts into a training set, 20% into a test set, and 10% into a validation set. Additionally, I ensured that the utterances of the speakers were approximately evenly distributed among the three sets. The number of files and the division between the subjects are shown in Table 3.

Finally, I also created a test set consisting of only typical speech data that matches the prompts and the number of files contained in the dysarthric test set. This set will be used to evaluate whether the performance of the recognition of typical speech data is affected by fine-tuning Whisper for dysarthric data.

Table 3: *Number of Utterances per Subject per Set*

| Subject | Training Set | Test Set | Validation Set |
|---|---|---|---|
| F03 | 752 | 205 | 112 |
| M03 | 561 | 148 | 89 |
| M02 | 544 | 140 | 80 |
| M01 | 530 | 128 | 78 |
| F04 | 480 | 126 | 72 |
| M04 | 461 | 114 | 62 |
| M05 | 324 | 90 | 47 |
| F01 | 158 | 48 | 22 |
| **Total** | **3810** | **999** | **563** |

### 3.2. Experiments

#### 3.2.1. *Zero-shot Testing (Baselines)*

Whisper is known for its capability in zero-shot transfer learning, where 'zero-shot' refers to the model's ability to recognise and adapt to new contexts it has not explicitly encountered during training, such as dysarthric speech [13]. Zero-shot testing therefore involves computing the WER achieved by Whisper before any fine-tuning is performed. This is necessary to establish the baseline performance of the various Whisper models on dysarthric speech.

To achieve this, all seven Whisper models established in Table 1 will be run against the dysarthric test set, and the WER will be calculated per model. Additionally, the same models will be tested on the typical speech test set to provide a comparative measure. These same steps will be performed on the models in the distil-Whisper family (Table 1). These baseline assessments will help determine the effectiveness of the Whisper models without any prior fine-tuning for dysarthric speech recognition and will determine which models will be used for fine-tuning.

#### 3.2.2. *Whisper Fine-Tuning*

Following the baseline assessment, the next step is to fine-tune the Whisper models using the TORGO training dataset. The fine-tuning procedure involves selecting the two best-performing Whisper models based on zero-shot testing results. Then, LoRA will be implemented using the PEFT library, with the PEFT parameters set to *r = 32*, *alpha = 64* and *dropout = 0.05*. The fine-tuning process involves training Whisper with a batch size of 32 and a learning rate of 1e-4. Additionally, 50 warmup steps and a linear decay are used. These values are all taken from [31], who performed a similar methodology for fine-tuning Whisper with child speech. In their study, these specific settings achieved a 38% reduction in relative WER using LoRA. This indicates that the chosen parameters are empirically validated and optimised for adapting Whisper to atypical speech data, particularly child speech. Similar to dysarthric speech, child speech tends to be more variable and challenging than typical speech. Furthermore, a temperature of 0.0 was used in the transcription of the audio files.

The number of epochs for which the models are trained will be determined using early stopping. Early stopping is a simple method of preventing overfitting to the training model. As explained in [32], early stopping involves training the model on a training set, but at the same time computing the error of the validation set at regular intervals. The training of the model should be stopped once the error on the validation set is higher than it was at the previous checkpoint.

#### 3.2.3. *Distil-Whisper Fine-Tuning*

In addition to fine-tuning the standard Whisper models, the distil-Whisper models will also undergo fine-tuning. To maintain comparability between the standard Whisper models and the distil-Whisper models, the distil-Whisper models will be chosen that correspond with the two standard Whisper models that are being fine-tuned. Furthermore, the same hyperparameters and early stopping will again be used.

## 4. Results

This section presents the results achieved via the aforementioned methodology. The results are divided into the three different experiments: the results for zero-shot testing are presented in Subsection 4.1, the fine-tuning of the standard Whisper models is shown in Subsection 4.2 and the fine-tuning results for the distil-Whisper models are shown in Subsection 4.3.

### 4.1. Zero-shot Testing

#### 4.1.1. *Standard Whisper Models*

In Table 4, the zero-shot results for the dysarthric test set across various Whisper models are presented. Detailed WERs for both typical and dysarthric test sets, categorised by dysarthria severity levels, are available in Table 14 in Appendix A. As anticipated, the WER generally decreases for dysarthric speech as the number of model parameters increases. The Large-V2 and Large-V3 models are the top performers, achieving WERs of 67.36% and 63.28% on the dysarthric test set, respectively.

Table 4: *Performance of Whisper Models on Dysarthric Test Set*

| Whisper Model | Average WER (%) |
|---|---|
| Tiny | 84.52% |
| Base | 76.36% |
| Small | 68.99% |
| Medium | 68.03% |
| Large-V1 | 74.51% |
| Large-V2 | 67.36% |
| Large-V3 | 63.28% |

It is interesting to note that the Large-V1 model performs worse than both the Small and Medium model, with WERs of 74.51% for the Large-V1 model, 68.03% for the Medium model and 68.99% for the Small model. This discrepancy is likely because the English-language Small and Medium models were used for transcription, whereas the Large-V1 model only has a multi-language model. Further analysis revealed that 9.9% of the Large-V1 transcriptions contained non-Roman characters, indicating non-English content. However, even though the Large-V2 and Large-V3 models also only contain a multilingual model, they outperform the English-only Medium and Small models. This suggests that training these models for more epochs and with more data has enabled them to better recognise dysarthric speech in zero-shot testing, even without knowing the language that is being spoken.

Figure 1 shows the WERs for the two best-performing models, Large-V2 and Large-V3, based on the severity of dysarthria. The results align with the expectation that WER increases with the severity of dysarthria, with the 'severe' dysarthria level consistently resulting in the highest WERs across both models. Notably, Large-V3 surpasses Large-V2 in recognising all severities of dysarthria, as well as typical speech.

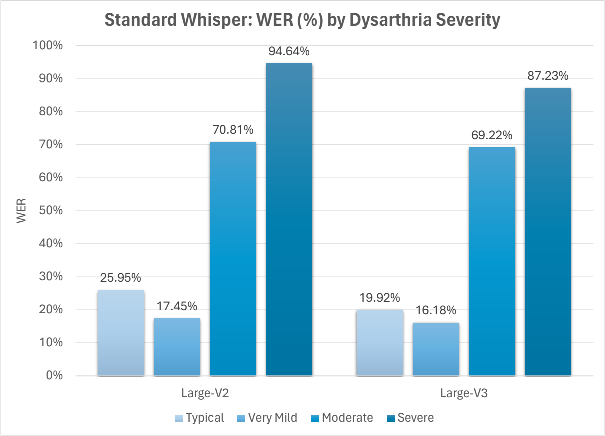

Figure 1: *Baseline WER (%) by Dysarthria Severity for Standard Whisper Models*

Interestingly, there is a significant difference in recognition performance between the 'very mild' and 'moderate' dysarthria severities. For both Large-V2 and Large-V3, the WER more than quadrupled from 'very mild' subjects to 'moderate' subjects. Additionally, it is noteworthy that 'very mild' dysarthric speech outperforms typical speech. Based on auditory inspection, one possible explanation is that whilst the typical and 'very mild' dysarthric speakers' pronunciations are similar, the dysarthric speakers are more intentional with their articulation and speak more slowly.

### *4.1.2. Distil-Whisper Models*

The full results of the zero-shot testing of all distil-Whisper models can be found in Table 15 in Appendix A. To allow for a comparison between the performance of the fine-tuned Whisper model and the fine-tuned distil-Whisper model, only the zero-shot performance of the distil-Large-V2 and distil-Large-V3 models will be evaluated.

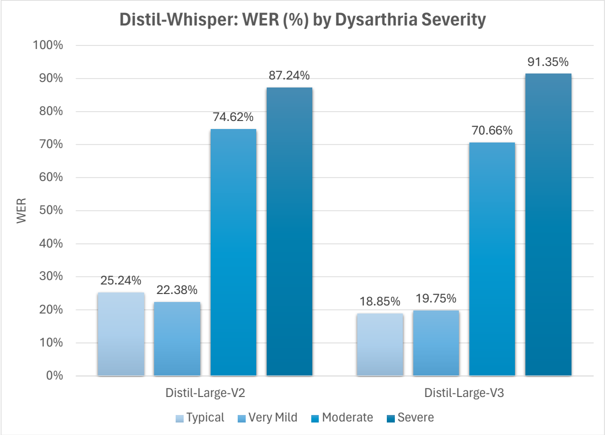

Figure 2: *Baseline WER (%) by Dysarthria Severity for Distil-Whisper Models*

Figure 2 shows the WER achieved by the Large-V2 and Large-V3 distil-Whisper models grouped by dysarthria severity. Similar to their standard counterparts, the 'severe' level displays the highest WERs, followed by the 'moderate' level. The lowest WERs are observed for typical and 'very mild' speech. Notably, the distil-Large-V2 model outperforms its non-distilled counterpart for 'severe' dysarthric speech, with 87.24% versus 94.64%.

## 4.2. Fine-tuning Whisper Models

The two standard Whisper models that best recognised dysarthric speech were the Large-V2 and Large-V3 models. Therefore, these two models will be used for fine-tuning. As shown for the Large-V2 model in Figure 3, the training and validation losses were tracked over 10 epochs. For this specific model, the training loss consistently decreased, since the model was increasingly fine-tuned to the training set. In contrast, the validation loss initially decreased rapidly but began to increase again after the first epoch. This suggests overfitting, so early stopping is used to determine that fine-tuning should conclude after one epoch.

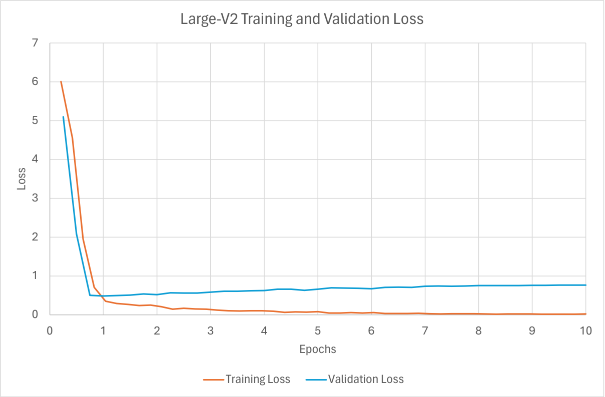

Figure 3: *Training and Validation Loss for Large-V2*

Table 5 shows the relative difference in WER after fine-tuning. The WER for 'severe' dysarthric speech has decreased by 21.78%, indicating a significant improvement in the model's ability to recognise speech from individuals with severe dysarthria. However, the WER for 'typical', 'very mild', and 'moderate' dysarthric speech has increased by 56.69%, 52.21%, and 20.77% respectively. This trend indicates that while the model's adjustments have benefited the recognition of severe dysarthric speech, these changes may have inadvertently degraded performance on less severe or typical speech patterns. One possible explanation for this is the trade-off between generalisation and specialisation that occurs when fine-tuning. Research indicates that fine-tuning models on specific datasets, such as dysarthric speech datasets, improve performance for those conditions but can reduce accuracy for standard or clean speech [33]. This is due to the model's shift in focus towards the new domain features, potentially neglecting the broader patterns learned during initial training.

Table 5: *Large-V2: Relative WER Difference*

| Speech Type | Relative WER |
|---|---|
| Typical | 56.69% |
| Very Mild | 52.21% |
| Moderate | 20.77% |
| Severe | -21.78% |

The Large-V3 model achieved its minimum validation loss after two epochs. The differences in WERs before and after fine-tuning are presented in Table 6. Notably, the WERs have drastically increased after fine-tuning. Upon closer inspection of the transcripts, it became evident that the fine-tuned model produced a significant number of so-called hallucinations.

Table 6: *Large-V3: Relative WER Difference*

| Speech Type | Relative WER |
|---|---|
| Typical | 1345.68% |
| Very Mild | 1179.73% |
| Moderate | 154.13% |
| Severe | 201.17% |

In the context of ASR, hallucinations are known as transcriptions created by an ASR model that are unrelated to the source audio [34]. Hallucinations are a known issue that can occur when using Whisper [35, 36]. The hallucinations in the fine-tuned Large-V3 model are severe, producing transcripts like "If you can do it can do it can do it can do it can do it can do it can do it can do it can do it can do it can do it can do it from the class, it, it, it, it, it, it". By counting the percentage of transcriptions containing words repeated three or more times [1], it was determined that 46.97% of the transcribed utterances of dysarthric speech and 24.72% of the transcribed utterances of healthy speech contained hallucinations. These hallucinations, typically characterised by the insertion of numerous additional words, significantly increased the overall WER.

Table 7: *Large-V3 after 1 Epoch: Relative WER Difference*

| Speech Type | Relative WER |
|---|---|
| Typical | -27.69% |
| Very Mild | -21.20% |
| Moderate | -41.32% |
| Severe | -24.53% |

Given the prevalence of hallucinations, it was hypothesised that the validation loss might not have accurately indicated the optimal number of training epochs, possibly due to the validation set not being a good representation of the test set. Consequently, the fine-tuned model was also evaluated after one epoch. The results are shown in Table 7. After fine-tuning the Large-V3 model for one epoch, all dysarthria severities exhibited decreased WERs compared to pre-fine-tuning, with the most significant decrease observed for 'moderate' dysarthric speech, which saw a reduction of 41.32%. Additionally, by re-evaluating the frequency of words occurring three or more times in a transcript, it was found that only 4.51% of the dysarthric and 3.70% of the typical transcriptions contained hallucinations. Overall, the Large-V3 model performs better after being fine-tuned for a single epoch, effectively recognising both typical speech and all severities of dysarthric speech. This contrasts the performance of the Large-V2 model, which performed worse for all classes of speech aside from the 'severe' dysarthric speech. This suggests that the larger amount of data that the Large-V3 model was trained on allows it to better generalise to a variety of dysarthria severity levels, in addition to typical speech.

### 4.3. Fine-tuning Distil-Whisper Model

The distil-Large-V2 model was trained for two epochs. As can be seen in Table 8, the 'very mild', 'moderate', and 'severe' dysarthric speech experienced reductions in relative WER by 7.55%, 14.54%, and 10.28% respectively. In contrast, the WER for typical speech increased after fine-tuning by a relative rate of 61.09%. The decrease in WERs for the dysarthric speech levels after fine-tuning indicates that the fine-tuning process improved the model's ability to recognise dysarthric speech.

However, the WER for typical speech increased because the fine-tuned model was trained exclusively on dysarthric data and not on additional typical speech data. As a result, the fine-tuned model is more specialised for dysarthric speech but performs worse on typical speech.

Table 8: *Distil-Large-V2: Relative WER Difference*

| Speech Type | Relative WER |
|---|---|
| Typical | 61.09% |
| Very Mild | -7.55% |
| Moderate | -14.54% |
| Severe | -10.28% |

The distil-Large-V3 model also experienced the lowest validation loss after two epochs. Table 9 shows the relative WER reduction before and after fine-tuning. The 'moderate' and 'severe' dysarthric speech experienced a reduction in relative WER by 20.49% and 17.54% respectively. In contrast, the WER for the typical and 'very mild' dysarthric speech increased after fine-tuning, by relative rates of 37.61% and 45.82%. Similar to the distilled Large-V2 model, this is probably because the fine-tuned model was only trained on dysarthric data, and not on further typical data. Therefore, the fine-tuned model has now become better tailored for dysarthric speech but performs worse on typical speech.

Table 9: *Distil-Large-V3: Relative WER Difference*

| Speech Type | Relative WER |
|---|---|
| Typical | 37.61% |
| Very Mild | 45.82% |
| Moderate | -20.49% |
| Severe | -17.54% |

## 5. Discussion

This section further analyses the achieved results by comparing the different model performances in Subsection 5.1. Furthermore, the achieved results will be discussed in the context of previous research in Subsection 5.2. Finally, Subsection 5.3 discusses some of the limitations of this research.

### 5.1. Comparison between Models

The relative WER differences for the various models across all dysarthria severities in the test set are shown in Table 10.

[1] This threshold was determined after manually inspecting the prompts and identifying that no prompts in the test set contain three or more repeated words.

The Large-V3 model (after fine-tuning it for one epoch) has achieved the largest decrease in relative WER for the entire dysarthric test set with 22.65%. Interestingly, the Large-V2 model achieved the lowest WER improvement with only 3.55%. The Large-V3 model is trained on significantly more speech data than the Large-V2 model, so this might indicate that models with larger training datasets are better at generalising to diverse and challenging speech patterns, such as those found in dysarthric speech.

Table 10: *Relative WER Difference for Different Models*

| Model | Relative WER Difference |
|---|---|
| Large-V2 | -3.55% |
| Large-V3 | -22.65% |
| Distil-Large-V2 | -11.44% |
| Distil-Large-V3 | -13.53% |

Furthermore, the distil-Large-V2 and distil-Large-V3 models, show moderate improvements of 11.44% and 13.53%, respectively. These results highlight that while model distillation can help in reducing the model size and inference time, there might be a trade-off in the extent of WER improvement achieved compared to the larger, non-distilled counterparts.

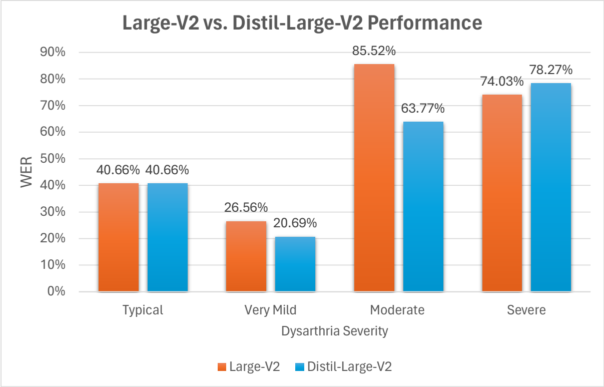

Figure 4: *Average WER after Fine-Tuning Large-V2 vs. Distil-Large-V2*

To make a more nuanced comparison between the performance of the standard Whisper models and the corresponding distilled models, Figure 4 and Figure 5 depict the average WER after fine-tuning per dysarthria severity level. For the Large-V2 models, the distilled version outperforms the standard models for all severities aside from 'severe' dysarthric speech. This does not match the hypothesis that the standard Whisper models will outperform the distilled models but matches the fact that the WERs increased for all severity levels aside from 'severe' dysarthria for the Large-V2 model (Table 5). This might be because the Large-V2 model is overfitting to the severe dysarthric speech, or it could be because the distilled models have a lower propensity for hallucinations than the standard Whisper model [18]. By counting the number of transcripts that contain words repeated three or more times, it can be seen that the non-distilled Large-V2 model produces hallucinations in 3.30% of its transcriptions, whilst the distil-Large-V2 model produced hallucinations 0% of the time.

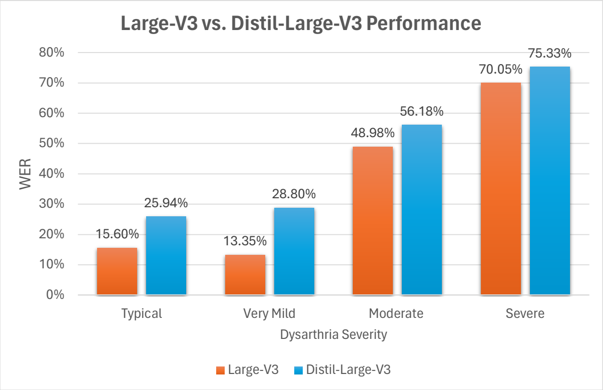

Figure 5: *Average WER after Fine-Tuning Large-V3 vs. Distil-Large-V3*

For the Large-V3 model, the standard version consistently outperforms the distilled version across all severity levels. This suggests that the larger number of parameters in the Large-V3 model allows it to better capture the features that make up and distinguish dysarthric from typical speech.

Nonetheless, the distilled models, while slightly less accurate, showcased improvements in runtime. Table 11 shows the time it took to fine-tune the models for 10 epochs. It can be seen that the distil-Large-V2 model took 21.93% less time to be fine-tuned than the standard Large-V2 and that the distilled Large-V3 model took 17.24% less time than the Large-V3 model. The time taken was relatively short across all models because a small training set and LoRA were used. However, when fine-tuning with a larger training set or for further epochs, the time efficiency of the distilled models will become especially advantageous. The distilled models are therefore particularly beneficial in scenarios with limited computational resources or the need for real-time processing. For the Large-V2 model, the distilled model outperformed the non-distilled model. However, for the Large-V3 model, a tradeoff between speed and performance will have to be made, depending on which model is chosen.

Table 11: *Time Taken for Fine-Tuning for 10 Epochs*

| Model | Time taken for Fine-Tuning |
|---|---|
| Large-V2 | 2 hours 15 minutes |
| Large-V3 | 2 hours 52 minutes |
| Distil-Large-V2 | 1 hour 46 minutes |
| Distil-Large-V3 | 2 hours 22 minutes |

### 5.2. Results Compared to Previous Literature

When comparing the results to previous studies, it is evident that transfer learning improves ASR performance for dysarthric speech. The performance of Whisper on dysarthric speech has consistently improved after fine-tuning, as supported by [12, 14, 15]. It is also insightful to examine the different models and datasets used in prior research and compare them to the results achieved in this study.

[15] used only Whisper's encoder in combination with a classifier, achieving a WER of 59.78% on the UA speech corpus. In contrast, this study used the full Whisper model and the TORGO dataset, achieving the lowest WER with the Large-V3 model at an average of 48.95% for the full dysarthric test set

(Table 21 in Appendix A). The lower WERs achieved in both studies despite differences in datasets and model components highlight the wide range of possibilities for improving Whisper to better recognise dysarthric speech.

[14] conducted a more similar experiment to the one presented in this paper, fine-tuning the Medium Whisper model with Dutch dysarthric speech data. This study achieved relative WER reductions of 7.3% for moderate dysarthria severity levels and 10.7% for severe dysarthria. In comparison, the highest improvements in this study were observed with the Large-V3 model, which saw relative WER reductions of 41.32% for moderate dysarthria and 24.53% for severe dysarthria. Although differing datasets and languages were used, these results suggest that models with more parameters (Large-V3 vs. Medium) are more effective in improving speech recognition performance through fine-tuning.

### 5.3. Limitations

Despite the progress demonstrated in this study, several challenges remain. The variability in dysarthric speech patterns continues to pose difficulties for ASR systems, as evidenced by the higher WERs in moderate and severe cases of dysarthria. Another limitation of this research is the exclusive use of the TORGO database, which primarily contains the spastic type of dysarthria [15]. Extending this work to include other types, such as flaccid, ataxic, and hypokinetic, would provide valuable insights into the generalisation of dysarthric speech recognition.

## 6. Conclusion and Future Work

This research aimed to improve automatic speech recognition for dysarthric speech by using transfer learning and Low-Rank Adaptation techniques on the Whisper model. The study compared the performance of distilled and non-distilled versions of the Whisper model, using the TORGO dataset for fine-tuning.

The results indicate that while the non-distilled Whisper model achieves higher accuracy in recognising dysarthric speech, the distilled model offers efficiency benefits, making it suitable for applications with limited computational resources. The Large-V2, Large-V3, distil-Large-V2 and distil-Large-V3 models all benefited from transfer learning, demonstrating substantial improvements in WER, thus aligning with previous research on ASR adaptation for atypical speech patterns. The largest performance improvement was achieved for the Large-V3 model with a relative WER reduction of 22.65% across dysarthria severities.

The results highlight the potential of advanced ASR technologies to improve accessibility for individuals with dysarthria, providing more accurate and reliable voice recognition interfaces. However, challenges such as the variability in dysarthric speech severity require further investigation to develop more robust solutions.

Based on the results of this research, several opportunities for future investigation have emerged. Firstly, while the WERs have decreased through fine-tuning, there remains potential for further improving the recognition of dysarthric speech. Future research could focus on using more extensive and diverse datasets for fine-tuning Whisper, as well as exploring other methods of transfer learning, such as deep transfer learning [37].

Additionally, investigating the fine-tuning of other state-of-the-art ASR models, such as wav2vec 2.0, would enable a comparative analysis of their performance and potential compared to the Whisper model. This could provide valuable insights into which models are best suited for improving ASR for dysarthric speech.

Finally, the occurrence of hallucinations when fine-tuning the Large-V3 model for two epochs warrants further investigation. Future research could explore the underlying causes of these hallucinations and develop strategies to mitigate them. Understanding and addressing this issue is crucial for enhancing the reliability and accuracy of fine-tuned ASR models.

## 7. Responsible Research

The Netherlands Code of Conduct for Research Integrity mentions specific standards for good research practice [38]. Keeping these in mind, Subsection 7.1 will examine how reproducible the presented research is. In Subsection 7.2, the responsible usage of data throughout the presented methodology will be discussed. Finally Subsection 7.3 examines how AI tools were used throughout the project.

### 7.1. Reproducibility

To ensure that the experiments described are reproducible and adhere to the FAIR principles (Findable, Accessible, Interoperable, Reusable) [39], the exact methodology is transparently reported:

- **Findable:** The only data used throughout this research was the TORGO dataset [22], which is a publicly available dataset. Furthermore, the codebase available on Gitlab[2] contains .txt files with the files that are part of the test, training and validation sets.
- **Accessible:** All the code used throughout this research is provided on Gitlab. Additionally, Subsection 3.2 explicitly states the hyperparameters used for the fine-tuning of both the Whisper model and the distil-Whisper models.
- **Interoperable:** The fine-tuned models are implemented using widely adopted frameworks such as PyTorch, which ensures they can be easily integrated into existing workflows for analysis, storage, and processing.
- **Reusable:** The fine-tuned models are publicly available on Hugging Face[3]. This allows the public to use these fine-tuned models, or to reuse them for further research.

### 7.2. Data Usage

All the data that was used to test on Whisper and fine-tune Whisper was taken from the TORGO dataset. This dataset is publicly available and was ethically sourced, as described in [22]. For example, no medical data is associated with the participants, aside from the cause of their dysarthria and a Frenchay assessment of motor skills. Furthermore, collecting the speech data involves working with vulnerable subjects, and this was considered throughout the data collection, e.g. by requiring that all subjects have a cognitive function at or above level VIII on the Rancho scale. Even though I did not collect this data myself, it is also important to consider the ethics behind data sourced by other researchers.

[2] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Zhang_Yue/mirellagunther-Exploring-state-of-the-art-speech-recognisers-for

[3] https://huggingface.co/mirella-guenther

### 7.3. Usage of LLMs

For complete transparency, ChatGPT and GitHub Copilot are tools that were used to improve writing or generate code. ChatGPT was primarily used to ensure that the report was written coherently and in a consistent writing style. This was achieved by providing a piece of text that I wrote, and then asking ChatGPT to "Please improve the coherence and flow of this text while sticking to the tone of a Bachelor thesis". However, I found that more often than not, ChatGPT provided very stilted writing, so I only ever took phrases or sentence structures from ChatGPT rather than using full paragraphs. ChatGPT was also a great aid for formatting tables, the bibliography and other figures in LaTeX. GitHub Copilot was used to aid in the writing of the code, usually by explaining what the next section of the code needed to do as a comment, and then using Copilot to generate the section.

## 8. References

[1] Leonard L LaPointe, Bruce E Murdoch, and Julie AG Stierwalt. *Brain-based communication disorders*. Plural Publishing, 2010.

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

[3] Robert DeRosier and Ruth S Farber. Speech recognition software as an assistive device: a pilot study of user satisfaction and psychosocial impact. *Work*, 25(2):125–134, 2005.

[4] Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[5] Mostafa Shahin, Usman Zafar, and Beena Ahmed. The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):400–412, 2020. doi:10.1109/JSTSP.2019.2959393.

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

[7] Susan Koch Fager and Judith M Burnfield. Speech recognition for environmental control: Effect of microphone type, dysarthria, and severity on recognition results. *Assistive Technology*, 27(4):199–207, 2015.

[8] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. In *Interspeech 2018*, page 471–475. ISCA, September 2018. URL: https://www.isca-archive.org/interspeech_2018/vachhani18_interspeech.html, doi:10.21437/Interspeech.2018-1751.

[9] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss. Simulating dysarthric speech for training data augmentation in clinical speech applications. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6009–6013. IEEE, 2018.

[10] Elizabeth Howarth, Geena Vabulas, Sean Connolly, Dawn Green, and Sara Smolley. Developing accessible speech technology with users with dysarthric speech. *Assistive Technology*, page 1–8, March 2024. doi:10.1080/10400435.2024.2328082.

[11] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[12] Yuki Takashima, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition. *IEEE Access*, 7:164320–164326, 2019.

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

[14] Lian Feenstra. *Aladdin, Alla dien, Allendien, please evaluate the performance of Whisper on Dutch dysarthric speech*. PhD thesis, Rijksuniversiteit Groningen, 2023.

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

[16] Heejin Kim, Mark Hasegawa Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang, Kenneth Watkin, Simone Frame, Harsh Vardhan Sharma, and Xi Zhou. Uaspeech, 2023. URL: https://dx.doi.org/10.21227/f9tc-ab45, doi:10.21227/f9tc-ab45.

[17] Whisper large-v3 - hugging face, Nov 2023. URL: https://huggingface.co/openai/whisper-large-v3.

[18] Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.

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

[20] S. Mangtulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul. Peft: State-of-the-art parameter-efficient fine-tuning, 2022. URL: https://github.com/huggingface/peft.

[21] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

[22] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, December 2012. doi:10.1007/s10579-011-9145-0.

[23] Kinfe Tadesse Mengistu and Frank Rudzicz. Adapting acoustic and lexical models to dysarthric speech. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4924–4927. IEEE, 2011.

[24] Frank Rudzicz. Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947–960, May 2011. doi:10.1109/TASL.2010.2072499.

[25] Cristina Espana-Bonet and José AR Fonollosa. 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*, pages 97–107. Springer, 2016.

[26] Enno Hermann and Mathew Magimai Doss. Dysarthric speech recognition with lattice-free mmi. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6109–6113. IEEE, 2020.

[27] Juliette Millet and Neil Zeghidour. Learning to detect dysarthria from raw speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5831–5835. IEEE, 2019.

[28] Z. Yue, H. Christensen, and J. Barker. Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition, October 2020. event-title: Interspeech 2020. URL: https://eprints.whiterose.ac.uk/164230/.

[29] Siddhartha Prakash. Deep learning-based detection of dysarthric speech disability, 2020.

[30] Zhengjun Yue, Feifei Xiong, Heidi Christensen, and Jon Barker. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6094–6098, Barcelona, Spain, May 2020. IEEE. URL: `https://ieeexplore.ieee.org/document/9054343/`, `doi:10.1109/ICASSP40776.2020.9054343`.

[31] Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D'Mello, and Jacob Whitehill. Automatic speech recognition tuned for child speech in the classroom. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12291–12295. IEEE, 2024.

[32] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[33] Katrin Tomanek, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim. On-device personalization of automatic speech recognition models for disordered speech. *arXiv preprint arXiv:2106.10259*, 2021.

[34] Rita Frieske and Bertram E Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*, 2024.

[35] Antonio Bevilacqua, Paolo Saviano, Alessandro Amirante, and Simon Pietro Romano. Whispy: Adapting stt whisper models to real-time environments. *arXiv preprint arXiv:2405.03484*, 2024.

[36] Eyal Liron Dolev, Clemens Fidel Lutz, and Noëmi Aepli. Does whisper understand swiss german? an automatic, qualitative, and human evaluation. *arXiv preprint arXiv:2404.19310*, 2024.

[37] Hamza Kheddar, Yassine Himeur, Somaya Al-Maadeed, Abbes Amira, and Faycal Bensaali. Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems*, 277:110851, 2023.

[38] KNAW, NFU, NWO, TO2-Federatie, Vereniging Hogescholen, and VSNU. Nederlandse gedragscode wetenschappelijke integriteit, 2018. URL: `https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:110600`, `doi:10.17026/DANS-2CJ-NVWU`.

[39] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

# A. Appendix A: Complete Results

This appendix presents the complete results obtained from the experiments conducted in this paper. Subsection A.1 contains the full results for the zero-shot testing experiment and Subsection A.2 contains the extensive results for the fine-tuning of both the standard and the distil-Whisper models.

## A.1. Complete Baseline Results

Table 12: *Average WER of Whisper Models on Typical and Dysarthric Speech*

| Whisper Model | Typical | Dysarthric |
|---|---|---|
| Tiny | 40.06% | 84.52% |
| Base | 43.38% | 76.36% |
| Small | 25.24% | 68.99% |
| Medium | 20.52% | 68.03% |
| Large-V1 | 24.10% | 74.51% |
| Large-V2 | 25.95% | 67.36% |
| Large-V3 | 19.92% | 63.28% |

Table 13: *Average WER of Distil-Whisper Models on Typical and Dysarthric Speech*

| Distil-Whisper Model | Typical | Dysarthric |
|---|---|---|
| Distil-Small | 27.54% | 68.99% |
| Distil-Medium | 21.75% | 63.25% |
| Distil-Large-V2 | 25.24% | 66.50% |
| Distil-Large-V3 | 18.85% | 66.46% |

Table 14: *Average WER of Whisper Models by Dysarthria Severity Level*

| Whisper Model | Typical | Very Mild | Moderate | Severe |
|---|---|---|---|---|
| Tiny | 40.06% | 34.63% | 91.39% | 109.51% |
| Base | 43.38% | 24.66% | 82.63% | 102.82% |
| Small | 25.35% | 18.02% | 72.93% | 98.03% |
| Medium | 20.52% | 14.71% | 73.10% | 96.24% |
| Large-V1 | 24.10% | 20.45% | 77.90% | 104.30% |
| Large-V2 | 25.95% | 17.45% | 70.81% | 94.64% |
| Large-V3 | 19.92% | 16.18% | 69.22% | 87.23% |

Table 15: *Average WER of Distil-Whisper Models by Dysarthria Severity Level*

| Distil-Whisper Model | Typical | Very Mild | Moderate | Severe |
|---|---|---|---|---|
| D-Small | 27.54% | 19.94% | 78.32% | 91.84% |
| D-Medium | 21.75% | 16.54% | 67.03% | 88.41% |
| D-Large-V2 | 25.24% | 22.38% | 74.62% | 87.24% |
| D-Large-V3 | 18.85% | 19.75% | 70.66% | 91.35% |

## A.2. Complete Fine-Tuning Results

Table 16: *Large-V2: Average WER Before and After Fine-Tuning*

| Severity Level | Before Fine-Tuning | After Fine-Tuning |
|---|---|---|
| Typical | 25.95% | 40.66% |
| Very Mild | 17.45% | 26.56% |
| Moderate | 70.81% | 85.52% |
| Severe | 94.64% | 74.03% |

Table 17: *Large-V3: Average WER Before and After Fine-Tuning*

| Severity Level | Before Fine-Tuning | After Fine-Tuning 1 Epoch | After Fine-Tuning 2 Epochs |
|---|---|---|---|
| Typical | 19.92% | 15.60% | 287.98% |
| Very Mild | 16.18% | 13.35% | 207.06% |
| Moderate | 69.22% | 48.98% | 175.91% |
| Severe | 87.23% | 70.05% | 262.71% |

Table 18: *Distil-Large-V2: Average WER Before and After Fine-Tuning*

| Severity Level | Before Fine-Tuning | After Fine-Tuning |
|---|---|---|
| Typical | 25.24% | 40.66% |
| Very Mild | 22.38% | 20.69% |
| Moderate | 74.62% | 63.77% |
| Severe | 87.24% | 78.27% |

Table 19: *Distil-Large-V3: Average WER Before and After Fine-Tuning*

| Severity Level | Before Fine-Tuning | After Fine-Tuning |
|---|---|---|
| Typical | 18.85% | 25.94% |
| Very Mild | 19.75% | 28.80% |
| Moderate | 70.66% | 56.18% |
| Severe | 91.35% | 75.33% |

Table 20: *Average WER on Entire Dysarthric Test Set*

| Model | Average WER |
|---|---|
| Large-V2 | 64.97% |
| Large-V3 | 48.95% |
| Distil-Large-V2 | 58.89% |
| Distil-Large-V3 | 57.47% |

Table 21: *Average WER on Entire Dysarthric Test Set*

| Model | Before Fine-Tuning | After Fine-Tuning |
|---|---|---|
| Large-V2 | 67.36% | 64.97% |
| Large-V3 | 63.28% | 48.95% |
| D-Large-V2 | 66.50% | 58.89% |
| D-Large-V3 | 66.46% | 57.47% |

## B. Appendix B: Original Zero-Shot Testing

Originally, zero-shot testing was conducted with a temperature setting of 1.0. This introduces an element of randomness, resulting in non-replicable outcomes. Despite this, the results have been included in this appendix to provide additional data. These results offer valuable insights into the performance of the models under slightly varied conditions, which may be beneficial for further analysis and comparison.

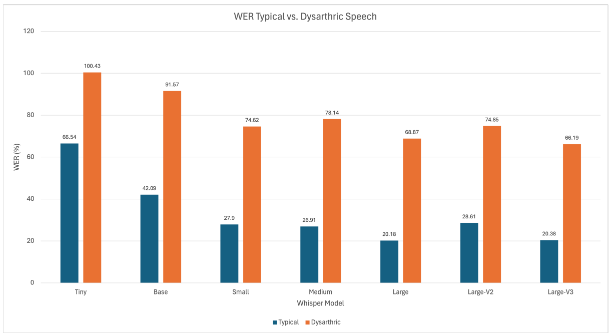

Figure 6: *Baseline WERs with Temperature 1.0*

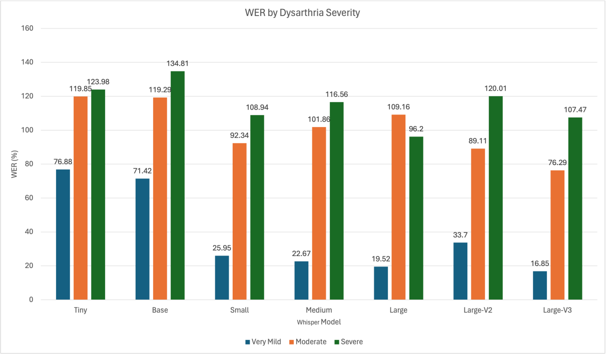

Figure 7: *Baseline WER by Dysarthria Severity with Temperature 1.0*

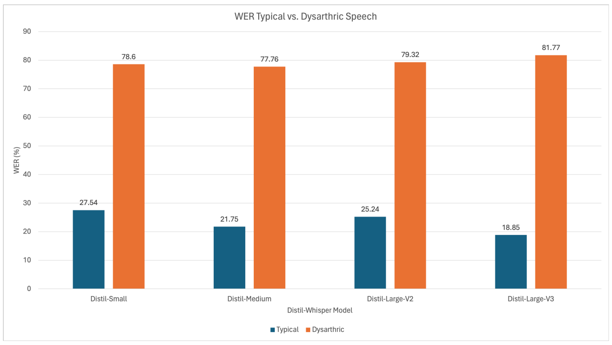

Figure 8: *Baseline WER per Distil-Whisper model with Temperature 1.0*