A Novel Deep Learning Based Speech Recognition System to Aid Communication for Dysarthria

[redacted]

Abstract

Dysarthria is a speech disorder that affects patients' lives negatively. Current Automatic Speech Recognition (ASR) systems handled for dysarthric patients are not able to properly transcribe their speech due to its poor articulations. Recently, there have been advancements in deep learning voice models that can be used to transcribe dysarthria speech to clear text, with one of the most promising being Whisper. This project aims in finetuning this model combined with the innovative methods of voice cloning and synthesis. This makes sure that the voices of dysarthric patients are heard through an improved ASR application that prompts users to input their own voice or choose a dysarthric file, translate the speech, and produce a clear voice output.

In order to find the most optimal model, different configurations of finetuning and testing with the UASpeech dataset and the TORGO dataset are used. Each dataset is given a preclassified severity scale for mild, moderate, and severe levels of disability, which allows for more comparisons to be made. For analysis, primary metrics like training loss, validation loss and Word Error Rate (WER) are used.

The ASR system developed in this research using the finetuned Whisper model demonstrates an improved transcription accuracy of approximately 70% for severe speech, 50% for moderate speech, and 54% for mild speech compared to baseline model. The framework of this system can pave way for future innovations by enabling better communication with dysarthric patients, creating personalized therapy, and facilitating early detection in speech decline.

Introduction

Dysarthria is a motor disorder of speech caused by abnormalities of the articulation and intelligibility of speech (Tomik & Guiloff, 2010), specifically in the strength, speed, range, tone, or accuracy of movements required for control of the respiratory and phonatory aspects of speech production (Pennington et al., 2013b). It is commonly associated with conditions such as Cerebral Palsy (CP) and Amyotrophic Lateral Sclerosis (ALS), affecting 40% and 80% of their patients respectively, while also being seen in other neurodegenerative disorders (Shih et al., 2022). For affected patients, a lot of problems tend to arise, such as weariness, depression, strained vocal quality, ataxic (loss of balance) symptoms, and spastic (related to muscle spasms) movement disorders (Schölderle et al., 2020).

The effects of dysarthria are rather inefficient in fluctuation in sounds, slurred or unclear sounds, lack of a proper rhythm, and it is challenging to understand the structure of words and sentences (Young & Mihailidis, 2010). The ASR systems that are already built cannot handle these speech distortions, especially for moderate to severely disabled patients (Young & Mihailidis, 2010).

Recent developments in machine learning and speech processing technologies provide promising answers to the problems of dysarthric voice transcription. OpenAI Whisper is a machine learning model that transcribes speech to text and it can be used in automatic speech recognition (ASR) as an effective strategy (Rathod et al., 2023). In dysarthric speech, the use of transfer learning significantly increases the accuracy of word recognition and transcription (Rathod et al., 2023).

Due to the lack of diverse datasets, various state of the art techniques for Text to speech were used. These included techniques such as voice cloning and speech synthesis. Voice cloning and synthesis techniques are powerful for text-to-speech conversion (Voice Cloning and Synthesis: Ultimate Guide, 2024). that was used to generate synthetic test dataset. Voice cloning is mimicking a specific individual voice to create a digital copy. On the other hand, voice synthesis generates natural sounding speech based on a textual input, which allows for more diverse models through synthetic voices (Chen et al., 2024).

Two well-known datasets of dysarthric patients are UASpeech and TORGO (Shih et al., 2022). The UASpeech dataset contains recordings from 15 individuals with dysarthria caused by CP and 13 controls. (Refer to Appendix 3 for the severity scale based on speech intelligibility) In contrast, the TORGO dataset has recordings from patients with CP and ALS, with a total of 9094 samples.

Word Error Rate (WER) was used to determine the model's accuracy. Series of experiments were conducted to chose the model that produces the least WER. Then, an interactive website is built to allow users select a dysarthric recording or provide a voice input, which is processed by the ASR model. The model then effectively converts the voice input into clear text that can be viewed and listened from the ASR system.

It is hypothesized that the finetuned Whisper model will outperform conventional models in transcribing speech to text for dysarthric individuals. The researcher aims this research can act as a tool to make sure the voices of dysarthric patients are heard. Thus, improving their communication in day to day life.

Materials

Coding Framework/Languages:

- Python, PyTorch, CSS, Gradio, Javascript, FastAPI, Draw.io, MS Excel

AI/ML Model:

- OpenAI Whisper, F5-TTS

Repository

- GitHub - Code Repository
- Hugging Face - Stores my fine-tuned model repository, Stores my preprocessed dataset
- WandB.AI - Stores results/graphs along with metrics like WER and Eval_Loss through tables

Server Details:

- Public Cloud: Vast.AI, AWS

Please refer to Appendix 2 for Server Details

General Methods

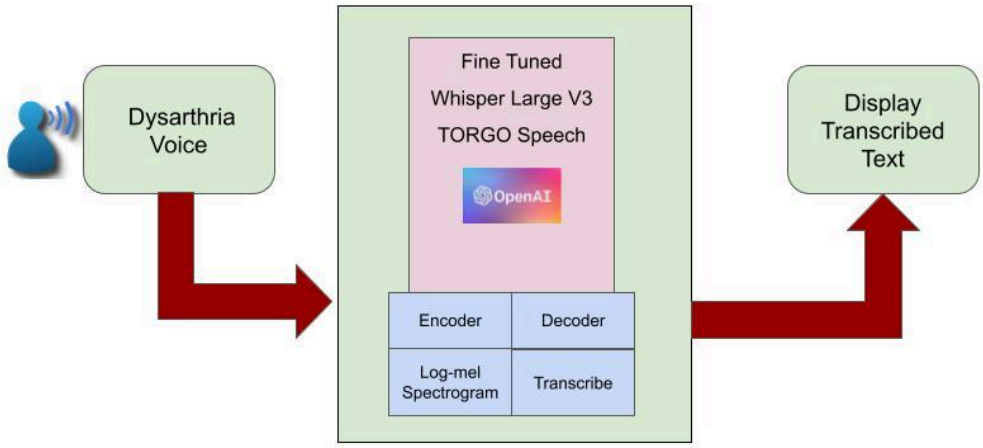
The ASR system, ClearVoiceAI, has 3 main components.

1. The input component receives an audio input.
2. The processing component analyze the audio input and transcribe them into text.
3. The output component display the transcribed text to the user along with the ability to read the text.

It was planned to make the system work in all mobile devices. Thus, a web based application was the optimal choice. Users will be able to record their speech and the ASR Engine will read and transcribe their speech to text, displaying it in the website.

Figure 1

Automatic Speech Recognition building blocks



Step 1: Selecting a Predefined Whisper V3 Large Model

Before building the website, the ASR Engine will need to transcribe the voice of dysarthria patients into clear text. A paper from the Google Research Team (Kolesnikov et al., 2019) showed that pre-trained models, when trained on large datasets, tend to have better visual representations. These visual representations when used for downstream tasks improved the scores on downstream tasks. Thus, it was decided to use a pre-trained model.

Then, research was done on the available ASR models that help to transcribe voice into text. An article by (Andrew Seagraves, 2022) reviewed Kaldi, Meta Wav2Vec 2.0 and OpenAI Whisper. In this article, the author compared these three models and tested them against various scenarios like spontaneous speech with noise and background speech, speech from phone calls, and multi person meeting calls with both in-person and video conferences. The article stated that the overall accuracy measured using the WER (this stands for the word error rate, which is the (number of substitutions + insertions + deletions)/(number of words in the sentence) all during transcription) is much better in Whisper compared to Wav2Vec 2.0 and Kaldi .

Whisper, which OpenAI released, revolutionized Robust Automatic Speech Recognition Systems (Radford et al., 2022) by training on a huge corpus of audio data, with the Whisper V3 Large model supporting 1.5 billion parameters with multilingual support (Sanchit Gandhi, 2022).

Moonshine is promising but research articles concluded that inputs with shorter length results in higher WER and there is no fine-tuning of the model released yet (Jeffries et al., 2024). The audio representations learned from these models, could lead to better finetuning.

When comparing the three models, Whisper's Word Error Rate (WER) was 45% less compared to Wav2Vec 2.0 and 63% less compared to Kladi. Therefore, Whisper V3 Large was utilized as the base pre-trained model.

Refer to Appendix 1.0 for model accuracy comparisons.

Step 2: Dataset Collection

Number of publicly available datasets were explored for dysarthric patients.. It was found that the TORGO dataset from University of Toronto (The University of Toronto, 2012) (Rudzicz et al., 2012) and UASpeech dataset from University of Illinois (Heejin Kim et al., 2023) are publicly available and it contains the speeches of dysarthric patients. It was found that the TORGO and UASpeech datasets had great pieces of samples for Whisper fine-tuning; they were also publicly available (Liu et al., 2024). The decision was further strengthened by Schu et al., 2022's article (Schu, Janbakhshi, et al., 2022), in which they received consistent and good accuracy from using ASR systems trained/tested on UASpeech and TORGO datasets. Since these datasets were collected in a controlled environment, there are biases in the dataset, which was referred to into that article. These datasets were classified into UASpeech, TORGO Speech, Voice Cloned from TORGO Speech, and Speech Synthesis created from TORGO Speech. Each classification included one-word and imperative utterances (imperative means that the prompt has ≥3 words in it).

UA Speech Dataset Collection:

The UASpeech dataset was used with ~57k single-word utterances (predominantly single-word) and the voices of 15 dysarthric patients and 13 controls. To create a well-balanced

dataset, classified UA Speech into five buckets based on a Severity Scale (A to E) - A being very mild and E being severe based on (Farhadipour & Veisi, 2023).

Refer to Appendix 3 for specifics on UASpeech Classifications/severity scales.

TORGO Dataset Collection:

5600 files with 4300 one-word utterances from the TORGO public database were used . The voices were from 8 dysarthric speakers (with CP and ALS) and 7 controls. Once the TORGO files were downloaded from the TORGO database, some preprocessing was done to split them into single and multiple utterances, with the files uploaded to Hugging Face. Note that a subset of the TORGO dataset is utilized (the 5600 used is not the full amount of samples TORGO offers) because it aligns with a predefined severity scale from (The University of Toronto, 2012).

Refer to Appendix 4 for TORGO Dataset Classifications/severity scales.

Refer to Appendix 5 for the TORGO Dataset Classification Process.

TORGO Voice Cloning Dataset Collection

Given the severe shortage of imperative sentences, the TORGO dataset (with multiple utterances) was augmented with voice cloned utterances, which are able to replicate the exact articulation of speech given. Publicly available advanced **text to speech (F5-TTS)** model based on deep learning technologies (Chen et al., 2024) was used. F5-TTS showed significant progress in the text to speech system, the models can also accurately voice clone.

Refer to Appendix 6 for the Voice Cloning Dataset Creation Process.

TORGO Speech Synthesis Dataset Collection

In speech synthesis, a Text-to-Speech (TTS) model is fine-tuned using the voices of dysarthric patients. Publicly available advanced **text to speech (F5-TTS)** model based on deep learning technologies (Chen et al., 2024) was used and the model was fine-tuned on the entire TORGO dataset. The biggest benefit with speech synthesis was no hallucinated voices were generated compared to voice cloning and thus no preprocessing was required.

Refer to Appendix 7 for the Speech Synthesis Dataset Creation Process.

The details of the datasets that includes UASpeech, TORGO Speech, Voice Cloned Speech and Speech Synthesis are summarized below in the following table:

Tota UA Speech	57000 files
Total Torgo (one word and imperative)	5600 files
Total Torgo One word Utterances	4300 files
Total Torgo Imperative	854 files
Total Torgo Cloned + Imperative	3870 files
Total Torgo Cloned + Synthesis	4270 files

Step 3: Finetune the pre trained Whisper Large V3 model

The pretrained publicly available Whisper Large v3 model was used as a base model to transcribe dysarthric speech into text. This model was fine tuned by utilizing different datasets, namely TORGO and UASpeech, and classifying each of their dysarthric patients in terms of their speech intelligibility. The main steps in finetuning are training and validation. Firstly, a part of the dataset will be chosen. In training, most of it will be used, and in validation, will use a smaller piece of it (usually around a 80% - 20% ratio between training and validation). The metrics used for fine tuning are training_loss, learning rate, validation loss or eval_loss, and WER (Word Error Rate) to check the quality of the fine tuned model, and then WER was used to check the accuracy during the testing process. Then, a model was tested using a completely different part of the dataset, and analyzed through metrics like WER. Note that all of the fine tuned models are approximately 5 GB in size.

Here are the general steps on the finetuning process started.

First, the fine-tuning code was developed (that includes sitting up training parameters based on this article (Sanchit Gandhi, 2022), in the development environment using PyCharm IDE. The code was then pushed and merged into Github.

Second, a training server (A6000 NVIDIA) on VastAI was created to train and fine tune the model. The researcher deployed the fine tuning code from Github to the training server and executed in the training server. After the model is finetuned model , it was uploaded to HuggingFace for future reference.

Third, a testing server in VastAI (RTX4090 NVIDIA) was created, which allowed for the deploying of the testing script from GitHub to the testing server on VastAI, which downloaded the fine-tuned model and dataset from HuggingFace to the testing server. Lastly, the testing script was executed and analyzed with metrics mentioned before for various dysarthric severity levels. The fine-tuned model with lowest WER will be used in my ClearVoiceAI (ASR) application as the default model.

The paper is broken up into experiment subsections. Here is a general overview of them (Note that when a "task" is mentioned, it consists of a few sub-experiments):

Firstly, the experiments were performed to see how the existing public, pretrained Whisper OpenAI worked for dysarthric speech. It was tested with UASpeech, which only has one word utterances and the TORGO dataset, filtered for both one word utterances and imperative sentences (multiple word utterances with words ≥ 3), separately.

Next, the Whisper large v3 model was fine tuned for dysarthric speech. It was trained using the UASpeech dataset that used shorter evaluation steps to make sure the model stopped overfitting after experimenting with a longer evaluation step. The UASpeech severity scale was used when picking datasets for each severity group, then the model was tested with this dataset as well as one word utterances of TORGO speech.

Then, the third major task was testing & analyzing previous fine tune shorter evaluation Model with different severity of the TORGO imperative sentence dataset.

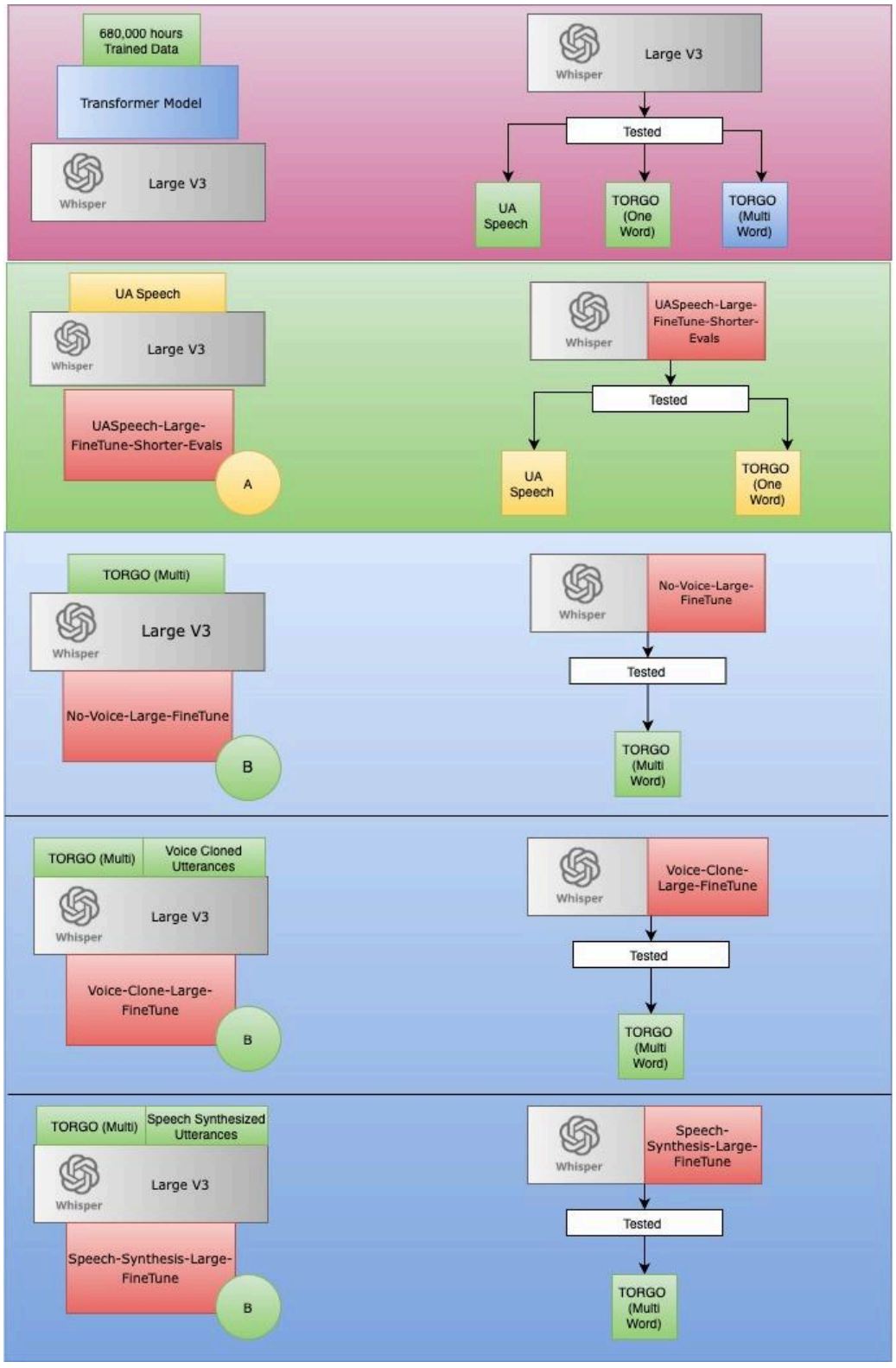
The fourth task was finetuning the large Whisper v3 model with a part of the imperative TORGO dataset and voice cloned utterances. It was then tested using a different severity of the imperative TORGO dataset.

The final task was finetuning the large Whisper v3 model with a part of the imperative TORGO dataset and speech-synthesized utterances. The model was tested on a different section on the TORGO imperative dataset just like before.

The image below represents all the experiments.

Figure 2

Fine-tuned Models Diagram

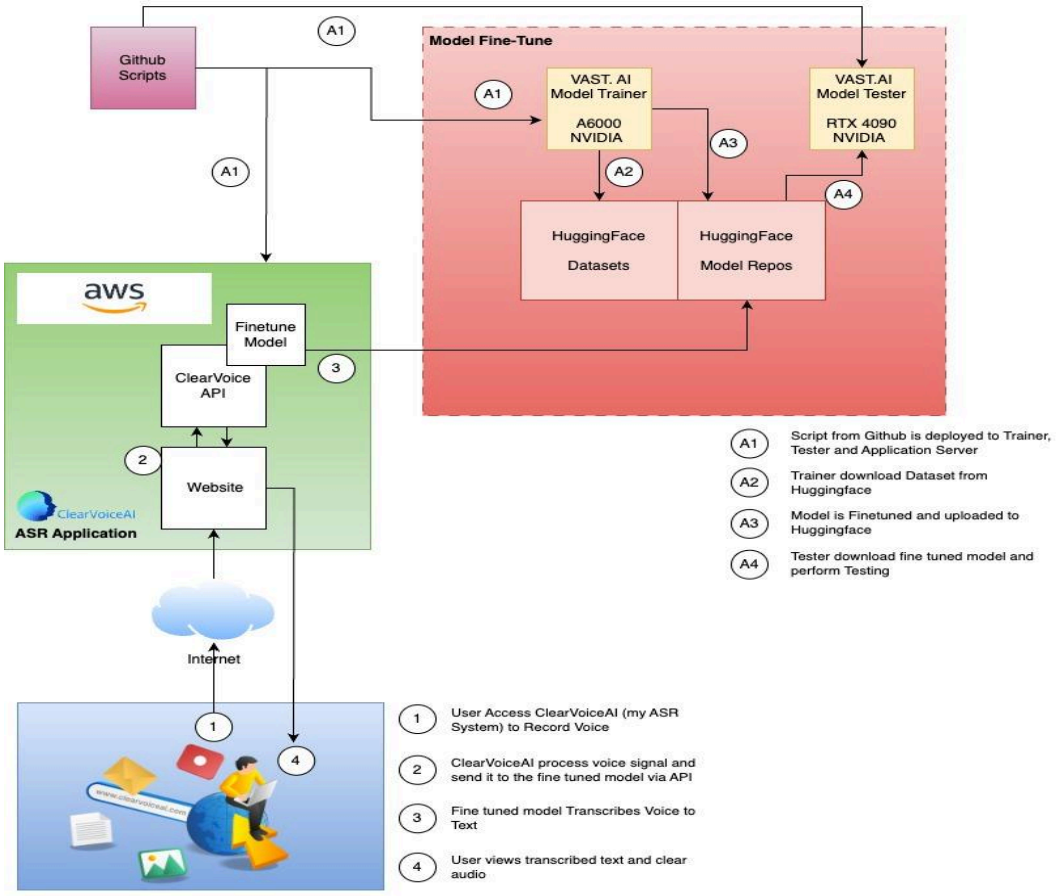


Step 4: Develop ASR Application

The ASR web application was named ClearVoiceAI. It was accessible from the URL, https://clearvoiceai.com. The application has front end layer was developed using Javascript, CSS and backend layer created using Python and FastAPI and fine tuned model trained using Python script. The application was hosted in AWS.

Figure 3

ASR Application (ClearVoice) System Diagram



1. When the user types https://clearvoiceai.com/, the browser renders the ASR web application.

2. Then, the user records their voice or chooses a pre-saved recording and then asks to transcribe it.
3. Now the request is sent to the API to process the audio file.
4. The first step in processing is to download the fine tuned model from the HuggingFace website.
5. Then the audio file is sent to the model, which then transcribes it to clear text .
6. This clear text is sent as response to the API, which in turn is sent to the website, and lastly to the browser.
7. Now the user can view the transcription along with a clear audio recording.

Experiment Methods and Results

The goal of the experiments is to fine tune OpenAI Whisper Large V3 model using UASpeech, TORGO publicly available and augmented datasets, created using Voice Cloning and Speech Synthesis methods. The experiment was started by understanding how to fine tune a Whisper model. The researcher learned from the article (Liu et al., 2024), when large models are fine tuned on smaller dataset tends to overfit quickly. The OpenAI-Whisper-Large-V3 model was selected as the base pre-trained model on which the researcher fine-tuned and created four fine-tuned models. 14 experiments were conducted to train, validate and test the four fine-tuned models.

Experiment 1 : Understanding how to Fine tune publicly available Whisper Small model

Methods

This experiment was performed to understand how to finetuned Whisper model. The researcher found an article in HuggingFace (Sanchit Gandhi, 2022) that explains how to fine tune Whisper Small model. This model utilized the Hindi datasets (Ardila et al., 2020) for training, validation and testing. To finetune the Whisper Small model, datasets were loaded, unwanted columns were removed. Then FeatureExtractor, Tokenizer and Processor were loaded and created. The data was prepared by converting all audio files to 16 kHZ, and training & evaluation parameters were set. The Word Error Rate (WER) was calculated after training the model. The model was trained for 2.3 hours based on the learning rate = 1e-5, maximum steps = 4000, and evaluation steps = 1000 .

Results

The model training loss reached zero at 4000 steps and validation loss of 0.4390. At this point, the model was no longer able to learn anything new. This point that the WER at 4000 steps would be the lowest/most optimal. It was calculated to be 32.4, meaning that 32.4% of the transcriptions were erroneous at some point, with a validation loss of 0.4390, meaning that when validating the model to the validation set, 43.90% of transcriptions were incorrect.

Refer to Appendix 8 for the result metrics and graphs.

Experiment 2: Testing and analyzing the Whisper V3 Large model with the UASpeech dataset

Methods

First experiments on Whisper Larger v3 model, started with UASpeech testing since it had predominantly one-word utterances, making it relatively simple to begin with. It also had ~57k datasets, which is a solid amount to finetune with. The Whisper Large V3 model is used for the remaining experiments. The severities that the dysarthric patients were classified from the dataset were based on a transcription accuracy metric. Groups were assigned from each of these categories for training and testing (it was around 60% –- 40% in terms of patient number for each group, but it followed the standard 80% - 20% in sample number for each group).

Here are the classifications of the patients in terms of speech intelligibility: 0-40% → Severe (2 patients: F03 and M12 with 5185 tested files), 40%-80% → Manageable (2 patients: M06 and M11 with 2847 tested files), >80% → Mild (2 patients: M08 and F05 with 10711 tested files).

Results (Appendix 9)

Overall, the WER for all scales was 127.57 (testing time was 1.45 hours total). For severe dysarthric speech, it was 142.57 (testing time was 31 minutes), for manageable dysarthric speech, it was 163.77 (testing time was 13 minutes), and for mild dysarthric speech, it was 110.68 (testing time was 50 minutes). The WER for this model was high for all severities with manageable being highest.

Experiment 3: Testing and analyzing the Whisper V3 Large model for the TORGO (one-word) dataset

Methods

TORGO has 4888 one-word utterances, not too much but still viable. It was divided into severe, manageable, and mild just like for the model that tested on UASpeech. For severe patients, groups M01, M02, M04, and F03 were used, with 2476 tested files. For manageable patients, group M05 was used, with 470 tested files. For mild patients, groups M03, F03, and F04 were used, with 1942 tested files. Note that this was all based on the severity scale(Appendix 5).

Results

Overall, the WER was 108.65 with a testing time of 21 minutes. For severe dysarthric speech, the WER was 122.95 with a testing time of 13 minutes, for manageable dysarthric speech, the WER was 114.71 with a testing time of 2 minutes, and lastly, for mild dysarthric speech, the WER was 86.24 with a testing time of 8 minutes. It can be seen that the WER was very high, for severe (A-E severity), 122.95 percent more words were erroneous than correct at some point of transcription, and it was 114.71 and 86.24 for manageable and mild respectively. This shows that this model still needs to be optimized, but it is slightly better compared to the previous model using the pretrained Whisper Large v3 and tested on UASpeech.

Refer to Appendix 10 for Dataset Classification, WER and Log Summary

Experiment 4: Testing and analyzing Whisper V3 Large model for TORGO speech and Tested with TORGO speech of various severities (multiple word utterances)

Methods

The patient groups were divided for TORGO a bit differently this time, with severe dysarthric speech containing M04 with 145 tested files, manageable containing M05 with 140 testing files, and mild containing F05 with 169 testing files. The total number of tested files is 454.

Results

For all scales, the overall results for WER was 43.17, with a tested time of 4 minutes. For severe severities, the WER was 90.95 with a tested time of 2 minutes. For manageable severities the WER was 34.08 with a tested time of 1 minute. For mild severities, the WER was 5.74 with a tested time of 1 minute. It can be seen that the WER was very high, for severe (A-E severity), 90.95 percent more words were erroneous than correct at some point of transcription, and it was 34.08 and 5.74 for manageable (C severity) and mild (A-B severity) respectively. This shows that this model still needs to be fine tuned.

Refer to Appendix 11 for Dataset Classification, WER and Log Summary

Experiment 5: Finetuning Whisper Large v3 trained on UASpeech and tested on UASpeech (Larger Evaluation)

Methods

In this experiment, the publicly available Whisper Large v3 model was fine-tuned by training and validating with UASpeech. For this experiment, max steps was set to 5000 and batch size was set to 16 . After running for 5000 steps with effective batch size of 16, the number of audio files it has seen = (5000*16), which translates to 2 epochs. The other processes were the same as the pretrained models. The experiment utilized total training files of 38.7k from 9 patients and total test files: 18.7k from 6 patients.

Refer to Table 3.2 UA Speech Training and Validation Dataset in Appendix 12 for dataset details. The model was trained for 6 hours.

Results

The training loss was zero around 4000 steps with validation loss of 0.3797 and at 4250 steps because the graph started overfitting (lowering its eval_loss to zero and then going up again after a certain amount of steps). Large models have a tendency to overfit (the training loss reduces but evaluation loss keeps increasing) .This means that the model is learning something specific to the dataset and this reduces generalizability. For example, if there is a glass dropping sound in the audio file and the utterance says “Baths”, the model might connect glass dropping sound with the word “Baths”. This is a crude example of overfit. The lowest evaluation loss was at around 1000 steps (approximate amount), so I took this in mind since its accuracy will be augmented if I continue with the amount of steps I had before.

Refer to Appendix 3 for Dataset UASpeech Classification

Refer to Appendix 12 for Training & Evaluation Result Table & Graph

Experiment 6: Finetuning Whisper Large v3 trained on UASpeech and validated on UASpeech (Shorter Evaluation)

Methods

The number of max_steps was reduced to 1500 just in case it overfitted slightly more than 1000. Eval_steps also got reduced → 100 due to the decrease in the max_steps. Model was trained for 6.5 hours with learning_rate = 1e-5, max_steps=1500 and eval_steps=100.

The experiment utilized total training files of 38.7k from 9 patients and total test files:18.7k from 6 patients. The classifications of patient groups were the same as in the Longer Eval model, now the evaluation steps is just reduced.

Results

Training loss approached zero (meaning the model was no longer able to learn anything new) with validation loss of 0.2762 was the closest at around 1500 steps. At the end, the WER was 28.41, meaning that 28.41% of the transcriptions were erroneous at some point, and the validation loss was 0.2762.

Refer to Appendix 3 for Dataset UASpeech Classification

Refer to Appendix 13 for Training & Evaluation Result Table & Graph

Experiment 7: Testing & Analyzing Large-finetune-shorter-evals Model and Test on UA Speech (one word utterance)

Methods

In this experiment, the large-finetune-shorter-evals model is tested with UASpeech of one word utterances. Therefore, the patient classifications (severe, manageable, mild, and patient groups) remain the same as the prior experiment but a subset of the validation dataset was used for testing and no training dataset. The number of files used for severe is 5185 (2 patients: F03 and M12) , for moderate is 2847 (2 patients: M06 and M11), and lastly, for mild, is 10711 (2 patients: M08 and F05).

Results

The WER from this experiment was 81.13 for severe severities with a tested time of 31 minutes, 18.34 for manageable severities with a tested time of 13 minutes, and lastly, 5.56 for mild severities with a tested time of 50 minutes. The researcher found that its WER was 43.1% less or better than it for the severe category, 88.8% less or better for the manageable category, and 95% less or better for the mild category of dysarthric speech comparing to the publicly available pre-trained model that was tested on UASpeech. Therefore the new model performed better transcription of all severities than the publicly available pre-trained model tested on UASpeech.

Refer to Appendix 14 for Dataset Classification, WER and Log Summary

Experiment 8 : Testing & Analyzing Large-finetune-shorter-evals Model and Test on TORGO Speech (one word utterance)

Methods

In this experiment, the Large-finetune-shorter-evals Model is tested on TORGO (one-word utterances). The patient classifications (severe, manageable, mild, and patient groups) remain the same, except now tested with TORGO speech . For severe, the researched used 2476 files (4 patients: M01, M02, M04, F03), for manageable, 470 files (1 patient: M05), and lastly, for mild, 1942 files (3 patients: M03, F03, F04).

Results

The WER for the transcription severe patients with this new model was 76.53 with a tested time of 13 minutes, 70.53 for manageable severities with a tested time of 2 minutes, and 49.87 for mild severities with a tested time of 8 minutes. Using the percent difference, we can see that the WER was 37.8% less/better for severe dysarthric speech, 38.5% less/better for manageable dysarthric speech, and lastly 42.2% less/better for mild dysarthric speech comparing this model to the publicly available pre-trained model that was tested on TORGO. Thus, the new model performed better on the transcription of all severities of dysarthric speech than the publicly available pre-trained model tested on TORGO.

Refer to Appendix 15 for Dataset Classification, WER and Log Summary

Experiment 9: Finetuning Whisper Large v3 trained and validated on TORGO Multiple Word Utterances

Methods

In this experiment, the researcher fine-tuned the Whisper Large V3 model with TORGO Imperative Sentences and validated it with imperative sentences (sentences with ≥3 words). To finetune the model, the datasets were loaded, unwanted columns were removed, loaded/created Whisper's FeatureExtractor + Tokenizer + Processor, prepared the data by converting all audio files to 16 kHZ, set training & evaluation parameters, trained the model , calculated WER and uploaded the model to Hugging Face website. The imperative sentences were filtered for training and testing. This dataset is smaller compared to UASpeech, thus 854 files were used for training with 5 patients ((M01,M02,F01,M03,F03), and 454 files (M04, M05, F04 → 3 patients) for validation. Model was trained for 1.5 hours with learning_rate = 1e-5, max_steps=1000 and eval_steps=100, one epoch ~ 53 steps with an effective batch size of 16. The model was trained for 1.5 hours.

Results

Training loss reached zero (meaning the model was no longer able to learn anything new) with validation loss of 0.4678 , most optimal stage at 1000 steps, after which the WER did not improve. At the end, the WER was 18.77, meaning that 18.77.% of the transcriptions were erroneous at some point, and the validation loss was 0.4678.

Refer to Appendix 16 for Dataset, Training & Evaluation Result Table & Graph

Experiment 10: Testing & Analyzing No-Voice-Clone-Large model and Test on TORGO Speech (multiple utterances)

Methods

In this experiment, the no voice clone large model was tested with TORGO speech with multiple word utterances. The dataset of 3 patients (M04, M05, F04) and 454 tested files, specifically 145 files for severe (M04), 140 files for manageable (M05), and 169 files for mild (F05) were used. Finally the WER tested against various severity dataset in this experiment was compared with the publicly available pre-trained model that is tested on TORGO (imperative sentences).

Results

The WER for the transcription severe patients with this new model was 33.04 with a tested time of 2 minutes, 17.52 for manageable severities with a tested time of 1 minute, and 2.5 for mild severities with a tested time of 1 minute. Using the percent difference, we can see that the WER was 63.7% less/better for severe dysarthric speech, 48.5% less/better for manageable dysarthric speech, and lastly 56.44% less/better for mild dysarthric speech comparing this model to the publicly available pre-trained model that was tested on TORGO (imperative sentences). Thus, the new model performed better on the transcription of all severities of dysarthric speech than the publicly available pre-trained model tested on TORGO (imperative sentences).

Refer to Appendix 17 for Dataset Classification, WER and Log Summary

Experiment 11: Finetuning Whisper Large v3 Trained with TORGO & Voice clone as data augmentation

Methods

In this experiment, the Whisper Large V3 model was trained and validated with TORGO & Voice clone dataset. Given the shortage of imperative sentences, the TORGO dataset (with multiple utterances) is augmented with voice cloned utterances. The details of voice cloning is summarized in the 'Voice Cloning' section (Appendix 6) in the data classification part. Removed any severe speeches that were garbled, this would mess up the voice cloning process. Total contains 4724 samples (voice cloning) and 854 imperative sentences, making ~5600 total samples of data in my new cloned dataset. For training, I used 5 patients, and for testing, I used 3 (M04, M05, F04). The model was trained for 3.5 hours.

Results

Training loss reached zero with validation loss of 0.4377 at 5000 steps. The WER was 15.36, meaning that 15.36% of the transcriptions were incorrect at some point. Ran this experiment for 5000 steps, since this was 5x more data .

Refer to Appendix 18 for Dataset, Training & Evaluation Result Table & Graph

Experiment 12: Testing & Analyzing Voice-Clone-Large model and Test on TORGO Speech (multiple utterances)

Methods

The testing dataset includes 454 files total, with 145 for severe (group M04), 140 for manageable (group M05), and 169 for mild (group F05). In this experiment, the researcher compared this model with the two previous ones with the TORGO dataset, namely the No-Voice-Clone Large model trained + tested on TORGO imperative sentences as well as the publicly available pre-trained model that is tested on TORGO.

Results

The WER for the transcription severe patients with this new model was 26.89 with a tested time of 2 minutes, 17.04 for manageable severities with a tested time of 1 minute, and 2.66for mild severities with a tested time of 1 minute. Using the percent difference, we can see that the WER was 70.04% less/better for severe dysarthric speech, 33.58% less/better for manageable dysarthric speech, and lastly 53.6% less/better for mild dysarthric speech comparing this model to the publicly available pre-trained model that was tested on TORGO . Thus, the new model performed better on the transcription of all severities of dysarthric speech than the publicly available pre-trained model tested on TORGO The WER for the transcription severe patients with this new model was 26.89 with a tested time of 2 minutes, 17.04 for manageable severities with a tested time of 1 minute, and 2.66 for mild severities with a tested time of 1 minute. Thus, the new model performed better on the transcription of all severities of dysarthric speech than the publicly available pre-trained model tested on TORGO.

Refer to Appendix 19 for Dataset Classification, WER and Log Summary

Experiment 13: Finetuning Whisper Large v3 Trained + Validated with TORGO & Speech Synthesis as data augmentation

Methods

In this experiment, the Whisper Large V3 model was trained and validated with TORGO & speech synthesis. Given the severe shortage of imperative sentences, the TORGO dataset (with multiple utterances) is augmented with speech synthesized dataset. The base F5-TTS model was finetuned which can effectively learn how to synthesize Dysarthic patient voices (Appendix 7). The details of speech synthesized is summarized in the 'Speech Synthesized section in the data classification part. No hallucinations (like voice cloning) and no manual preprocessing was required. When we use voice cloning, we are not using single word utterances. With this approach, we are finetuning the base F5-TTS model on the entire TORGO dataset, which might generate much better data. The model was trained on 5 patients with 5124 files for 3.5 hours.

Results

Training loss reached zero with validation loss of 0.4259 at 5000 steps. The WER was 16.8396, meaning that 16.8396% of the transcriptions were incorrect at some point. Ran this experiment for 5000 steps, since this was 5x more data. The model's WER reduces drastically up to ~4200 steps and more or less plateaus after that.

Refer to Appendix 20 for Dataset, Training & Evaluation Result Table & Graph

Experiment 14 : Testing & Analyzing Speech-Synth-Large-Finetune model and Test on TORGO Speech (multiple utterances)

Methods

In this experiment, the speech synthesis model was tested with TORGO imperative sentences. The testing dataset includes 454 total tested files and patient groups M04 → severe (145 files), M05 → manageable (140 files), and F05 → mild (169 files). Researcher also compared this speech synthesis model to the voice cloning model.

Results

The WER for the transcription severe patients with this new model was 29.44 with a tested time of 4 minutes, 18.49 for manageable severities with a tested time of 1 minute, and 3 for mild severities with a tested time of 1 minute. Using the percent difference, we can see that the WER was 9.5% less/better for severe dysarthric speech, 8.5% less/better for manageable dysarthric speech, and lastly 12.8% less/better for mild dysarthric speech comparing this model to the publicly available pre-trained model that was tested on TORGO . Thus, the new model performed better on the transcription of all severities of dysarthric speech than the publicly available pre-trained model tested on TORGO.

Refer to Appendix 21 for Dataset Classification, WER and Log Summary

Discussion

Conclusion

This study's objective was to build a functional model that would effectively turn dysarthric speech into text. The results indicate that the Voice-clone-large-fine tuned Whisper v3 model outperformed the public available pre-trained Whisper model from OpenAI in transcribing dysarthria speech into text , supporting my hypothesis. Specifically, the WER was 70% lower for severe patients, 50% lower for moderate severity patients, and 54% lower for mild patients. The results show a promising future for ASR systems when voice cloning technique is incorporated to generate more synthetic dataset. This makes the model to be robust with better accuracy.

ClearVoiceAI, an ASR application was built to support and apply the model in real life. The ClearVoice API, website, and fine tuned model were hosted in AWS. The code was saved on Github, then sent to training and testing servers for finetuning. The finetuned model was uploaded to HuggingFace. The application allowed the users to choose a sample dysarthric speech or record their own voice. It then produced a transcribed voice output.

The study had few drawbacks as they did not have many samples for TORGO and UASpeech datasets. The total amount of imperative sentences used was only 4728, which is low for a Whisper-based model. Also, environmental factors like background noises would affect the quality of speech heard in the application. So, performance of the ASR system would not be optimal in the real world. Lastly, the website was not tested for scale to handle a large number of users.

Future Work

The future study plans on making the application act as a diagnostic tool that can be used to monitor the progression of dysarthric symptoms. This plan graphs WER changes over time to help caregivers track treatment progress. An increasing WER indicates that dysarthria symptoms are not improving. Thus allowing caregivers to change their intensity of speech treatment.

If possible, the research is then planned to also be able to create an improved model that can transcribe not just dysarthria, but also the speech of other disorders such as dysphagia, apraxia, and aphasia, which will also be included into the diagnostic tool mentioned above.

Based on how the diagnostic utility performs, a feature that provides recommendations to each patient may be integrated to make sure the right treatment is given to them by their doctors, nurses, and/or caregivers.

References

Andrew Seagraves. (2022, December 19). 3 Best Open-Source ASR Models Compared: Whisper, wav2vec 2.0, Kaldi – Insights & Usability | Deepgram. Deepgram. https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 4211–4215.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2024, October 9). F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. arXiv.org. https://arxiv.org/abs/2410.06885

Diana Nguyen. (2022a). *ngdiana/uaspeech_severity_low · Datasets at Hugging Face*. Hugging Face. https://huggingface.co/datasets/ngdiana/uaspeech_severity_low

Diana Nguyen. (2022a). *ngdiana/uaspeech_severity_high · Datasets at Hugging Face*. Hugging Face. https://huggingface.co/datasets/ngdiana/uaspeech_severity_high

Farhadipour, A., & Veisi, H. (2023). Gammatonegram representation for End-to-End dysarthric speech processing tasks: speech recognition, speaker identification, and intelligibility assessment. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.03296

Heejin Kim, Mark Hasegawa Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang, Kenneth Watkin, Simone Frame, Harsh Vardhan Sharma, Xi Zhou. (2023). UASpeech. IEEE Dataport. https://dx.doi.org/10.21227/f9tc-ab45

Jeffries, N., King, E., Kudlur, M., Nicholson, G., Wang, J., & Warden, P. (2024, October 21). *Moonshine: Speech recognition for live transcription and voice commands*. arXiv.org. https://arxiv.org/abs/2410.15608

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019, December 24). Big Transfer (BIT): General Visual Representation Learning. arXiv.org. https://arxiv.org/abs/1912.11370

Lettergram. (2019). *sentence-classification/data/imperatives.csv at master · lettergram/sentence-classification*. GitHub. https://github.com/lettergram/sentence-classification/blob/master/data/imperatives.csv

Liu, Y., Yang, X. & Qu, D. Exploration of Whisper fine-tuning strategies for low-resource ASR. J AUDIO SPEECH MUSIC PROC. 2024, 29 (2024). https://doi.org/10.1186/s13636-024-00349-3

Pennington, L., Roelant, E., Thompson, V., Robson, S., Steen, N., & Miller, N. (2013b). Intensive dysarthria therapy for younger children with cerebral palsy. Developmental Medicine & Child Neurology, 55(5), 464–471. https://doi.org/10.1111/dmcn.12098

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December 6). *Robust speech recognition via Large-Scale Weak Supervision*. arXiv.org. https://arxiv.org/abs/2212.04356

Rathod, B., et al. (2023). Transfer learning using Whisper for dysarthric automatic speech recognition. In *Proceedings of the International Conference on Speech Technologies* (pp. 419–431). Springer. https://doi.org/10.1007/978-3-031-48309-7_46

Rudzicz, F., Namasivayam, A.K. & Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resources & Evaluation 46, 523–541 (2012). https://doi.org/10.1007/s10579-011-9145-0

Sanchit Gandhi. (2022, November 3). *Fine-Tune Whisper for Multilingual ASR with Transformers*. https://huggingface.co/blog/fine-tune-whisper

Schölderle, T., Haas, E., & Ziegler, W. (2020). Dysarthria syndromes in children with cerebral palsy. *Developmental Medicine & Child Neurology*, *63*(4), 444–449. https://doi.org/10.1111/dmcn.14679

Schu, G., Janbakhshi, P., & Kodrasi, I. (2022, November 16). *On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches*. arXiv.org. https://arxiv.org/abs/2211.08833

Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., & Shih, M.-H. (2022). Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit. *Healthcare*, *10*(10), 1956. https://doi.org/10.3390/healthcare10101956

SWivid. (n.d.). *GitHub - SWivid/F5-TTS: Official code for "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching."* GitHub. https://github.com/SWivid/F5-TTS/tree/main

The University of Toronto. (2012). *The TORGO Database: Acoustic and Articulatory Speech from speakers with Dysarthria*. https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis*, *11*(1–2), 4–15. https://doi.org/10.3109/17482960802379004

Voice Cloning and Synthesis: Ultimate guide. (2024). Farasoft. https://www.forasoft.com/blog/article/voice-cloning-synthesis

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. https://doi.org/10.1080/10400435.2010.483646

Appendix

Appendix 1

Table 1.1

Model Training Summary

Model	**Training hours**	**Training Method**
Whisper	700,000	Supervised Training
Wav2Vec2	60,000	Unsupervised Training

Credit: (Andrew Seagraves, 2022)

Table 1.2

Model Accuracy using Overall WER

Dataset	**Kaldi**	**wav2vec 2.0**	**Whisper**
Conversational AI	64.2	36.3	19.9
Phone call	69.9	31.0	16.6
Meeting	44.0	27.4	13.9
Earnings Call	65.8	28.1	9.7
Video	47.6	23.3	8.9

Credit: (Andrew Seagraves, 2022)

Appendix 2

Table 2.1: Server details

	Host	**Server**	**VRAM Memory (GB)**	**Storage (GB)**	**Hourly Rate ($)**	**GPU Type**
Training	Vast AI	A6000	48	250	0.8	NVIDIA A6000 (1 GPU, 48GB memory)
Testing	Vast AI	RTX 4090	24	250	0.421	NVIDIA 4090 (1 GPU, 24 GB memory)
App	AWS	g6.2xlarge	24	450	1.3456	NVIDIA L4 (1 GPU, 24 GB GPU Memory)

Appendix 3

Table 3.1: UA Speech Dataset

Total Patient	15
Total UA Speech files	57000
Low Severity UA Speech files	34900
High Severity UA Speech files	22500

Credit: (Diana Nguyen, 2022) and (Heejin Kim et al., 2023)

Figure 3.1

UA Speech Dataset Classification based on Severity. *Credit: (Farhadipour and Veisi, 2023)*

No.	Speaker ID	gender	Age	Speech Intelligibility
1	F02	Female	30	29%
2	F03	Female	51	6%
3	F04	Female	18	62%
4	F05	Female	22	95%
5	M01	Male	>18	15%
6	M04	Male	>18	2%
7	M05	Male	21	58%
8	M06	Male	18	39%
9	M07	Male	58	28%
10	M08	Male	28	93%
11	M09	Male	18	86%
12	M10	Male	21	93%
13	M11	Male	48	62%
14	M12	Male	19	7.4%
15	M14	Male	40	90.4%
16	M16	Male	>18	43%

Table 3.2

UA Speech Training and Validation Dataset

Severity Scale	Severity Type	Total	Training (9 patients , 38.7k files)	Validation (6 patients, 18.7k files)
0-40%	Severe	6 patient F02, F03, M01, M04, M07, M12	4 patient F02, M01, M04, M07	2 Patient F03, M12
40-80%	Manageable	4 Patient F04, M05, M06, M11	2 Patient F04, M05	2 Patient M06, M11
> 80%	Mild	5 Patient M14, M10, M09, M08, F05	3 Patient M14, M10, M09	2 Patient M08, F05

Appendix 4

Table 4.1

TORGO Dataset Classification

Total Patient	8
Control	7
Total TORGO Speech files Used	5600

Total Size	18.5 GB
Dysarthria Female Patient	F01, F03, F04
Dysarthria Female Controls (only used for feeding proper prompts)	FC01, FC02, FC03
Dysarthria Male Patient	M01, M02, M03, M04, M05
Dysarthria Male Controls (only used for feeding proper prompts)	MC01, MC02, MC03, MC04

Credit: (The University of Toronto, 2012) and (Rudzicz et al., 2012)

Appendix 5

TORGO Dataset Classification Process

The dataset when downloaded has notes, prompts folder under each patient and their session. Under the notes folder, there is a {patient_id}.csv, and it has a row called Intelligibility . Select the value from Sentence level Intelligibility and manually noted down. Only single word utterances are chosen for comparison. Classify dataset based on Severity Scale (A to E) - A being very mild and E being SEVERE.

Figure 5.1

Sample - M05.csv.

	In Speech	c/d
Tongue	At Rest	c
	Protrusion	c
	Elevation	e
	Lateral	e
	Alternate	d
	In Speech	c/d
Intel.	Words	a
	Sentences	c
	Conversation	c
Summary	When M05 gets emotional	jis speech system (breath
	Soeech->slow	laborious; words/utterances can be understood when speaking slowly. He did need
Session1		
Bite-plate normalization trial (head correction) : #3		
Coil 4 switched for coil 2		

Credit: (*The University of Toronto, 2012) and (Rudzicz et al., 2012)*

Appendix 6

Voice Cloning Dataset Creation Process

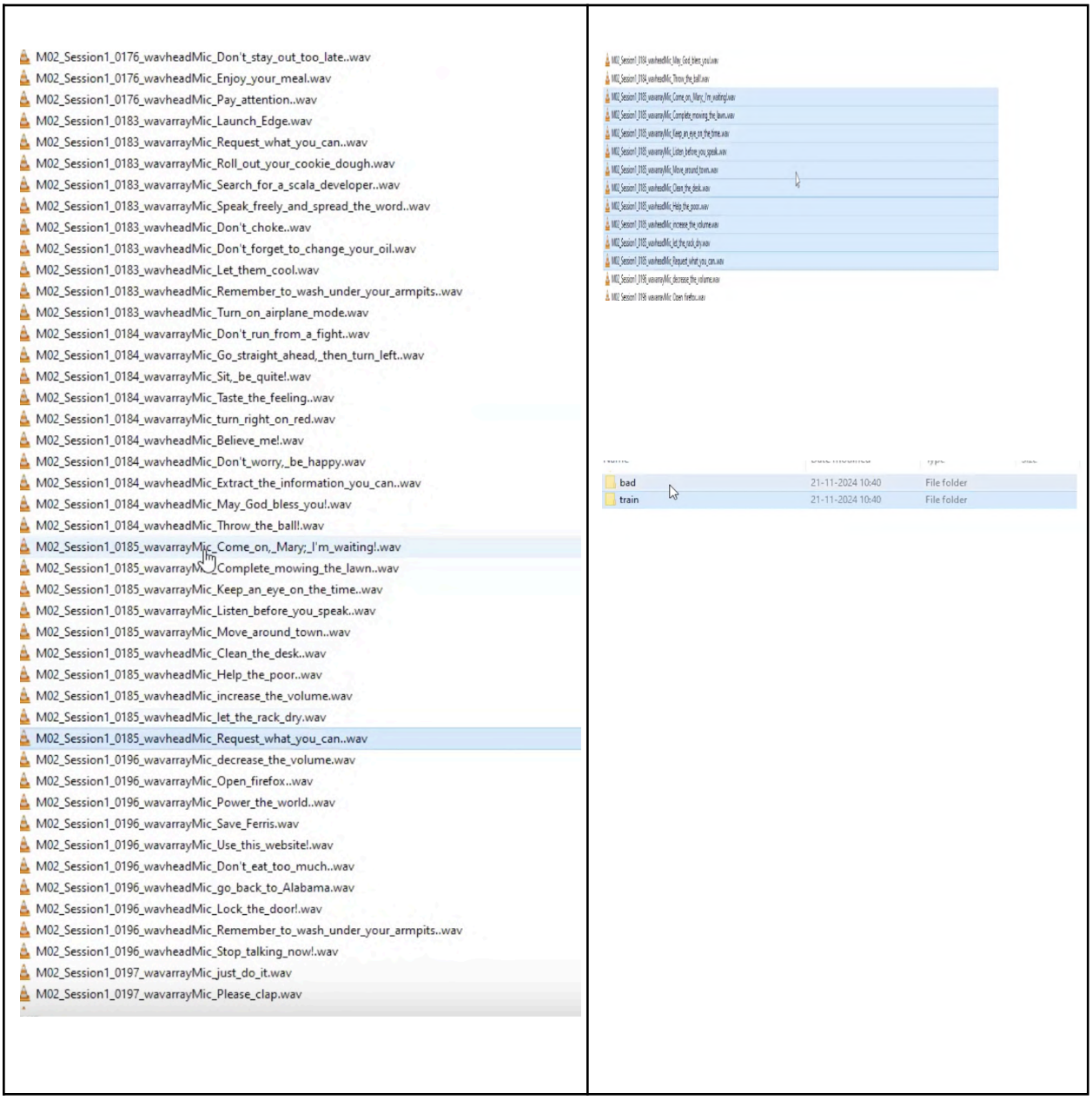
Publicly available advanced **text to speech (F5-TTS)** model based on deep learning technologies (Chen et al., 2024) was used. F5-TTS showed significant progress in the text to speech system, the models can also accurately voice clone. The researcher wanted to generate imperative sentences that related to real world scenarios. Imperative sentences from lettergram (Lettergram, 2019). Using the F5-TTS method, the researcher was able to generate 704 Imperative sentences from Lettergram to decide what dysarthria audio file to generate. Using audio from Training dataset patients as reference and with sample 5 random statements , voice cloned wav files were generated.

As a next step, the hallucinated or not valid audio files were manually removed. Voice cloned wav file is only as good as the reference audio. If the reference audio is very garbled, the

generated audio will hallucinate. In the below example (Figure 6.1) , there are 10 files (that include wavarray and wavhead). Wavarray and Wavhead location of the patient/mic when recording. The researcher listened to the synthetic voices of patients and if one of the folder sentence is bad, he discarded all files in the wrong. It took him one week to manually clean the data and 185 hallucinated files were discarded.

Figure 6.1

Folder with Voice Cloned files



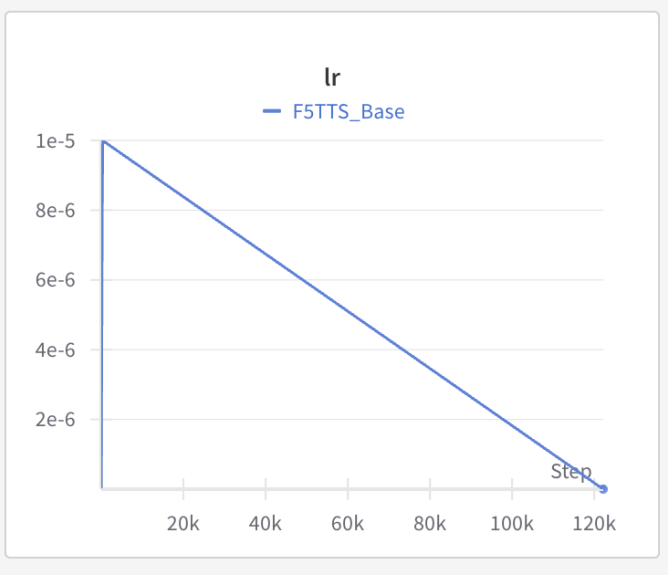
Appendix 7

Speech Synthesis Dataset Creation Process

Publicly available advanced **text to speech (F5-TTS)** model based on deep learning technologies (Chen et al., 2024) was used. Base F5-tts model was finetuned which can effectively learn how to synthesis Dysarthic patients voices.Model trained with training ran overnight (10 hours) ~120k steps. No Hallucinations (like voice clones) . No preprocessing was required for this process.

Figure 7.1

Training and learning rate



Appendix 8

Understand, Analyze, Fine tune publicly available Whisper Small model

Table 8.1

Experiment 1 Result table

Model	Trained On	Tested On	WER	Validation Loss	Trained Time
OpenAI Whisper small	Common voice - Mozilla - Hindi - Train +Validation	Common voice - Mozilla - Hindi - Test	32.4	0.4121	2.3 hours

Credit: (Ardila et al., 2010)

Figure 10.1

Experiment 1 Training & Evaluation Configuration

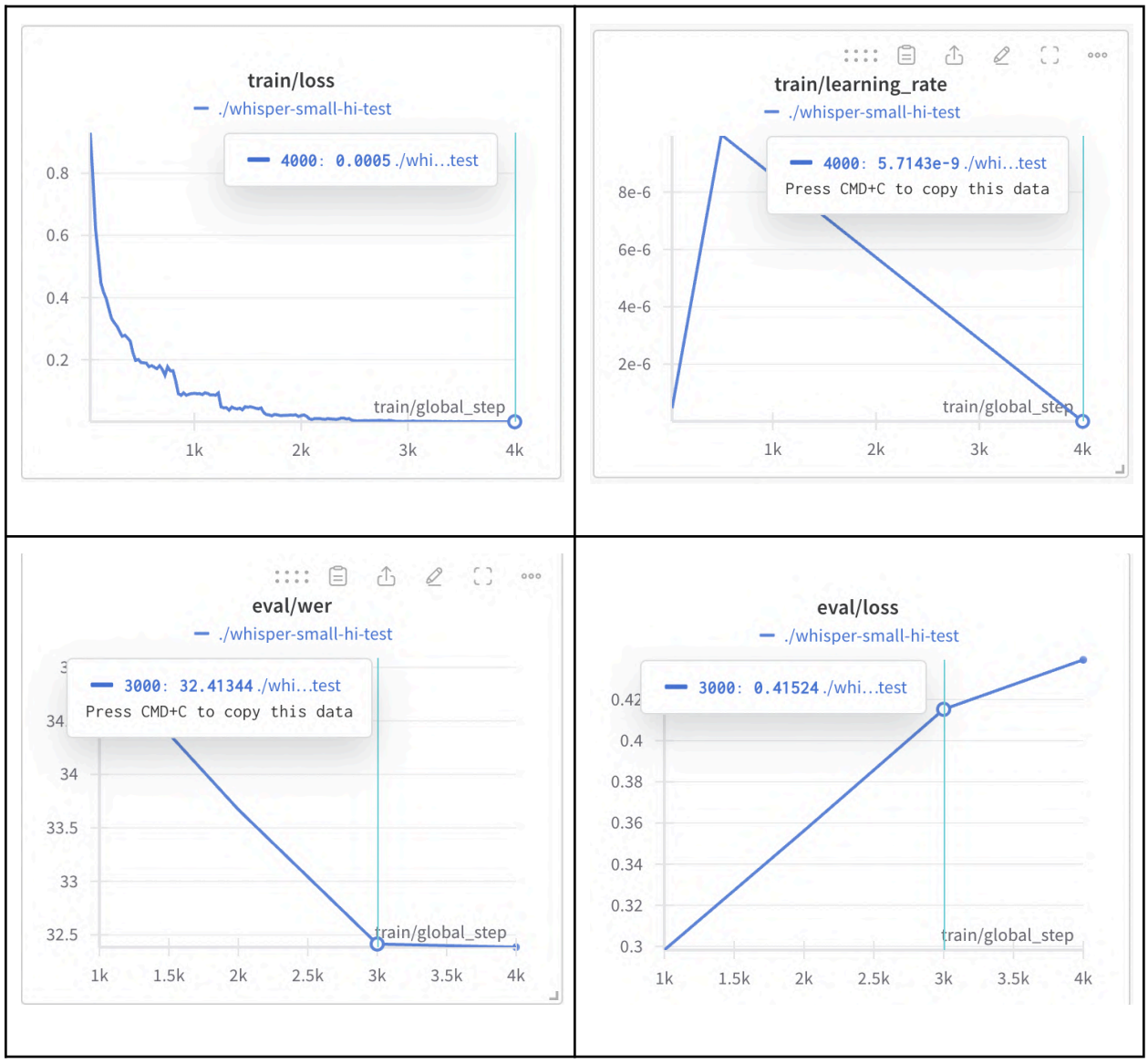
Training results

Training Loss	Epoch	Step	Validation Loss	Wer
0.0922	2.4450	1000	0.2977	35.0038
0.0209	4.8900	2000	0.3548	34.0430
0.0013	7.3350	3000	0.4121	32.3584
0.0004	9.7800	4000	0.4390	32.4854

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./whisper-small-hi-test",
    per_device_train_batch_size=16,
    gradient_accumulation_steps=1,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=4000,
    gradient_checkpointing=True,
    fp16=True,
    evaluation_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=1000,
    eval_steps=1000,
    logging_steps=25,
    report_to=["tensorboard"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
)
```

Figure 10.2

Experiment 1 Training & Evaluation Result



Appendix 9

Test & Analyze Whisper V3 Large model for UA speech with UA Speech of various severities

Table 9.1

Experiment 2 Dataset

Severity Scale	Severity Type	Testing (6 patients total)	No of Files
0-40%	Severe	2 Patient (F03, M12)	5185
40-80%	Manageable	2 Patient (M06, M11)	2847
> 80%	Mild	2 Patient (M08, F05)	10711

Table 9.2

Experiment 2 WER (Word Error Rate) Result Table

Severity Scale	WER	Tested Time
All Scale	127.57	1.45 hours
Severe	142.57	31 minutes
Manageable	163.77	13 minutes
Mild	110.68	50 minutes

Table 9.2

Experiment 2 Logs-1

All Scale	2024-11-28 07:35:48,997 - INFO - Processing batch 1101 2024-11-28 07:38:45,750 - INFO - Processing batch 1126 2024-11-28 07:41:09,236 - INFO - Processing batch 1151 2024-11-28 07:43:37,622 - INFO - Word Error Rate (WER) for model openai/whisper-large-v3: 127.57% 2024-11-28 07:43:37,640 - INFO - CSV saved for model openai/whisper-large-v3 at: predictions_and_ground_truth_openai_whisper-l 2024-11-28 07:43:37,984 - INFO - Cleared GPU cache for model openai/whisper-large-v3
Severe	2024-11-28 12:09:45,840 - INFO - Processing batch 251 for severity: Severe 2024-11-28 12:12:22,204 - INFO - Processing batch 276 for severity: Severe 2024-11-28 12:14:39,167 - INFO - Processing batch 301 for severity: Severe 2024-11-28 12:17:06,793 - INFO - WER for model openai/whisper-large-v3 and severity Severe: 142.57% 2024-11-28 12:17:06,801 - INFO - CSV saved for model openai/whisper-large-v3 and severity Severe at: predicti 2024-11-28 12:17:06,802 - INFO - Processing severity level: Manageable with model: openai/whisper-large-v3
Manageable	2024-11-28 12:25:00,464 - INFO - Processing batch 101 for severity: Manageable 2024-11-28 12:26:53,376 - INFO - Processing batch 126 for severity: Manageable 2024-11-28 12:28:43,913 - INFO - Processing batch 151 for severity: Manageable 2024-11-28 12:30:33,318 - INFO - Processing batch 176 for severity: Manageable 2024-11-28 12:30:50,318 - INFO - WER for model openai/whisper-large-v3 and severity Manageable: 163.77% 2024-11-28 12:30:50,326 - INFO - CSV saved for model openai/whisper-large-v3 and severity Manageable at: pre(

Table 9.3

Experiment 2 Logs-2

Mild	2024-11-28 13:14:56,574 - INFO - Processing batch 601 for severity: Mild 2024-11-28 13:16:42,551 - INFO - Processing batch 626 for severity: Mild 2024-11-28 13:18:27,425 - INFO - Processing batch 651 for severity: Mild 2024-11-28 13:20:07,422 - INFO - WER for model openai/whisper-large-v3 and severity Mild: 110.68% 2024-11-28 13:20:07,433 - INFO - CSV saved for model openai/whisper-large-v3 and severity Mild at: predictio 2024-11-28 13:20:07,751 - INFO - Cleared GPU cache for model openai/whisper-large-v3

Appendix 10

Test & Analyze Whisper V3 Large model for UA speech and Tested with TORGO (one word utterances)

Table 10.1

Experiment 3 Dataset

Severity Type	Testing	No Of Tested Files
Severe	M01, M02, M04, F03	2476
Manageable	M05	470
Mild	M03, F03, F04	1942

Table 10.2

Experiment 3 WER (Word Error Rate) Result Table

Severity Scale	Public Openai Whisper - Tested on TORGO	Tested Time
All Scale	108.65	21 min
Severe	122.95	13 min
Manageable	114.71	2 min
Mild	86.24	8 min

Table 10.3

Experiment 3 Logs-1

All Scale	2024-11-28 10:48:08,096 - INFO - Processing batch 263 2024-11-28 10:48:12,209 - INFO - Processing batch 264 2024-11-28 10:48:16,326 - INFO - Processing batch 265 2024-11-28 10:48:20,497 - INFO - Processing batch 266 2024-11-28 10:48:30,107 - INFO - Word Error Rate (WER) for model openai/whisper-large-v3: 108.65% 2024-11-28 10:48:30,114 - INFO - CSV saved for model openai/whisper-large-v3 at: predictions_and_ground_truth_ 2024-11-28 10:48:30,316 - INFO - Cleared GPU cache for model openai/whisper-large-v3
Severe	2024-11-28 15:13:56,886 - INFO - Processing batch 141 for severity: SEVERE 2024-11-28 15:14:17,820 - INFO - Processing batch 146 for severity: SEVERE 2024-11-28 15:14:38,478 - INFO - Processing batch 151 for severity: SEVERE 2024-11-28 15:15:00,882 - INFO - WER for model openai/whisper-large-v3 and severity SEVERE: 122.95% 2024-11-28 15:15:00,887 - INFO - CSV saved for model openai/whisper-large-v3 and severity SEVERE at: predictior
Manageable	2024-11-28 15:15:52,822 - INFO - Processing batch 11 for severity: MANAGABLE 2024-11-28 15:16:13,863 - INFO - Processing batch 16 for severity: MANAGABLE 2024-11-28 15:16:53,526 - INFO - Processing batch 21 for severity: MANAGABLE 2024-11-28 15:17:20,583 - INFO - Processing batch 26 for severity: MANAGABLE 2024-11-28 15:17:39,153 - INFO - WER for model openai/whisper-large-v3 and severity MANAGABLE: 114.71% 2024-11-28 15:17:39,156 - INFO - CSV saved for model openai/whisper-large-v3 and severity MANAGABLE at: predictions_and_ground_
Mild	2024-11-28 15:24:51,416 - INFO - Processing batch 106 for severity: MILD 2024-11-28 15:25:11,646 - INFO - Processing batch 111 for severity: MILD 2024-11-28 15:25:31,761 - INFO - Processing batch 116 for severity: MILD 2024-11-28 15:25:51,754 - INFO - Processing batch 121 for severity: MILD 2024-11-28 15:25:59,714 - INFO - WER for model openai/whisper-large-v3 and severity MILD: 86.24% 2024-11-28 15:25:59,718 - INFO - CSV saved for model openai/whisper-large-v3 and severity MILD at: predictions_and_grou

Appendix 11

Test & Analyze Whisper V3 Large model for TORGO speech and Tested with TORGO speech of various severities (multiple word utterances)

Table 11.1

Experiment 4 Dataset

Severity Type	No of Tested Patient	No Of Tested Files
Severe	M04	145
Manageable	M05	140
Mild	F05	169

Table 11.2

Experiment 4 WER (Word Error Rate) Result Table

Severity Scale	Public Openai Whisper - Tested on TORGO (multiple word utterances)	Tested Time
All Scale	43.17	4 min
Severe	90.95	2 min
Manageable	34.08	1 min
Mild	5.74	1 min

Table 11.2

Experiment 4 Logs

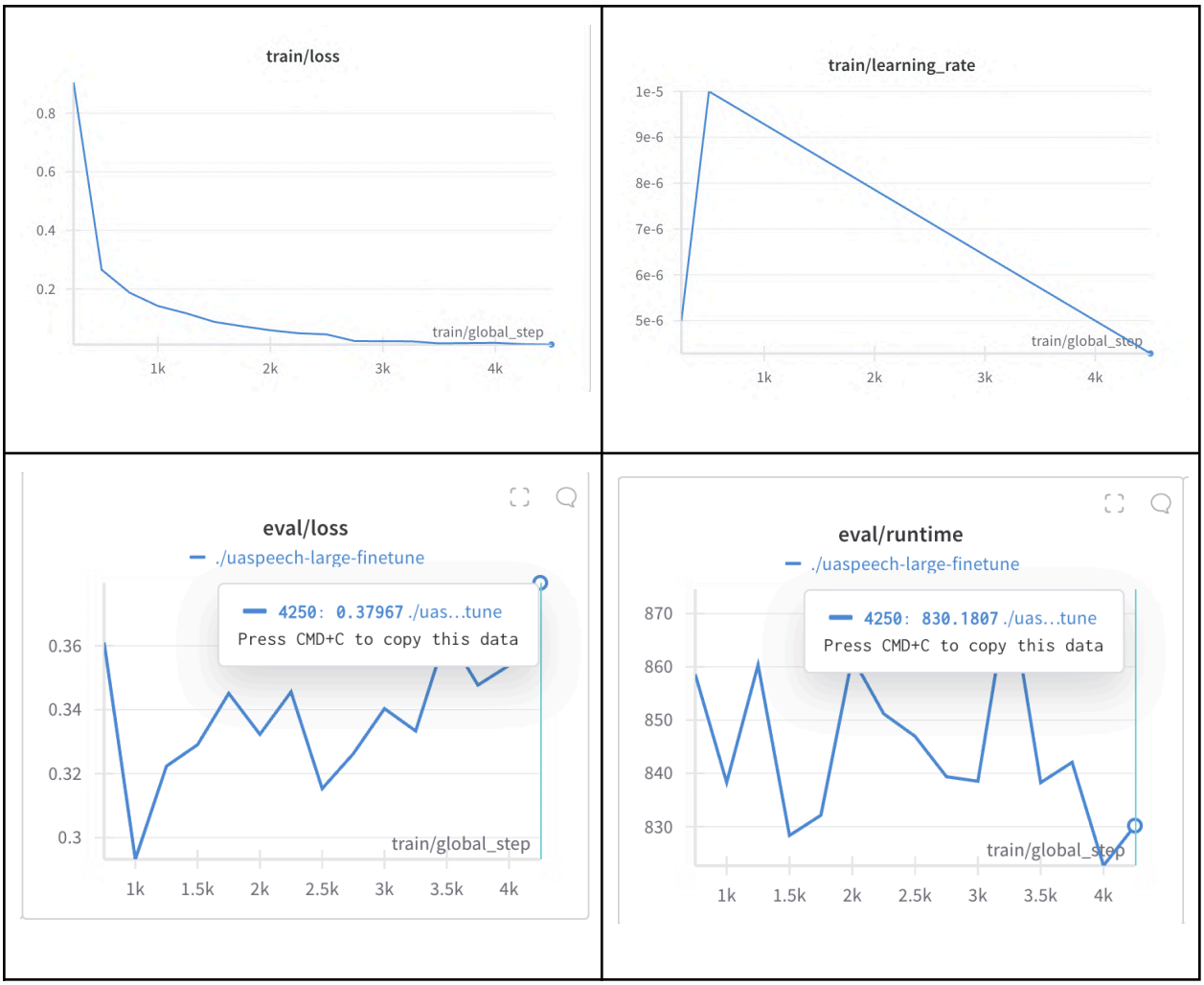
All Scale	2024-11-28 17:05:03,561 - INFO - Processing batch 26 2024-11-28 17:05:10,300 - INFO - Processing batch 27 2024-11-28 17:05:17,444 - INFO - Processing batch 28 2024-11-28 17:05:24,520 - INFO - Processing batch 29 2024-11-28 17:05:31,939 - INFO - Word Error Rate (WER) for model openai/whisper-large-v3: 43.17% 2024-11-28 17:05:31,943 - INFO - CSV saved for model openai/whisper-large-v3 at: predictions_and_ground_truth_openai_whisper-large-v3_torgo.csv 2024-11-28 17:05:32,144 - INFO - Cleared GPU cache for model openai/whisper-large-v3
Severe	2024-11-29 00:56:37,372 - INFO - Processing batch 6 for severity: Severe 2024-11-29 00:56:43,791 - INFO - Processing batch 7 for severity: Severe 2024-11-29 00:56:50,852 - INFO - Processing batch 8 for severity: Severe 2024-11-29 00:56:57,438 - INFO - Processing batch 9 for severity: Severe 2024-11-29 00:57:03,587 - INFO - Processing batch 10 for severity: Severe 2024-11-29 00:57:04,967 - INFO - WER for model openai/whisper-large-v3 and severity Severe: 90.95% 2024-11-29 00:57:04,970 - INFO - CSV saved for model openai/whisper-large-v3 and severity Severe at: predictions_and_ground_truth_imperative_openai_whisper-large-v3_Severe.csv
Manageable	2024-11-29 00:57:31,300 - INFO - Processing batch 5 for severity: Manageable 2024-11-29 00:57:37,389 - INFO - Processing batch 6 for severity: Manageable 2024-11-29 00:57:43,367 - INFO - Processing batch 7 for severity: Manageable 2024-11-29 00:57:49,353 - INFO - Processing batch 8 for severity: Manageable 2024-11-29 00:57:55,332 - INFO - Processing batch 9 for severity: Manageable 2024-11-29 00:58:00,931 - INFO - WER for model openai/whisper-large-v3 and severity Manageable: 34.08% 2024-11-29 00:58:00,933 - INFO - CSV saved for model openai/whisper-large-v3 and severity Manageable at: predictions_and_ground_truth_imperative_openai_whis
Mild	2024-11-29 00:58:43,771 - INFO - Processing batch 8 for severity: Mild 2024-11-29 00:58:50,495 - INFO - Processing batch 9 for severity: Mild 2024-11-29 00:58:57,445 - INFO - Processing batch 10 for severity: Mild 2024-11-29 00:59:04,092 - INFO - Processing batch 11 for severity: Mild 2024-11-29 00:59:09,798 - INFO - WER for model openai/whisper-large-v3 and severity Mild: 5.74% 2024-11-29 00:59:09,800 - INFO - CSV saved for model openai/whisper-large-v3 and severity Mild at: predictions_and_ground_truth_imperative_op

Appendix 12

Finetune Whisper Large v3 trained on UASpeech and tested on UASpeech (Larger Evaluation)

Graph 12.1

Experiment 5 Training & Evaluation Result



Appendix 13

Finetune Whisper Large v3 trained on UASpeech and tested on UASpeech (Shorter Evaluation)

Table 13.1

Table 13.1: Experiment 6 Result table

Model	Trained On	Tested On	WER	Validation Loss	Trained Time
Large-finetune-shorter-evals Model	UASpeech	UASpeech (one word utterances)	28.41	0.2762	6.5 hours

Figure 13.1

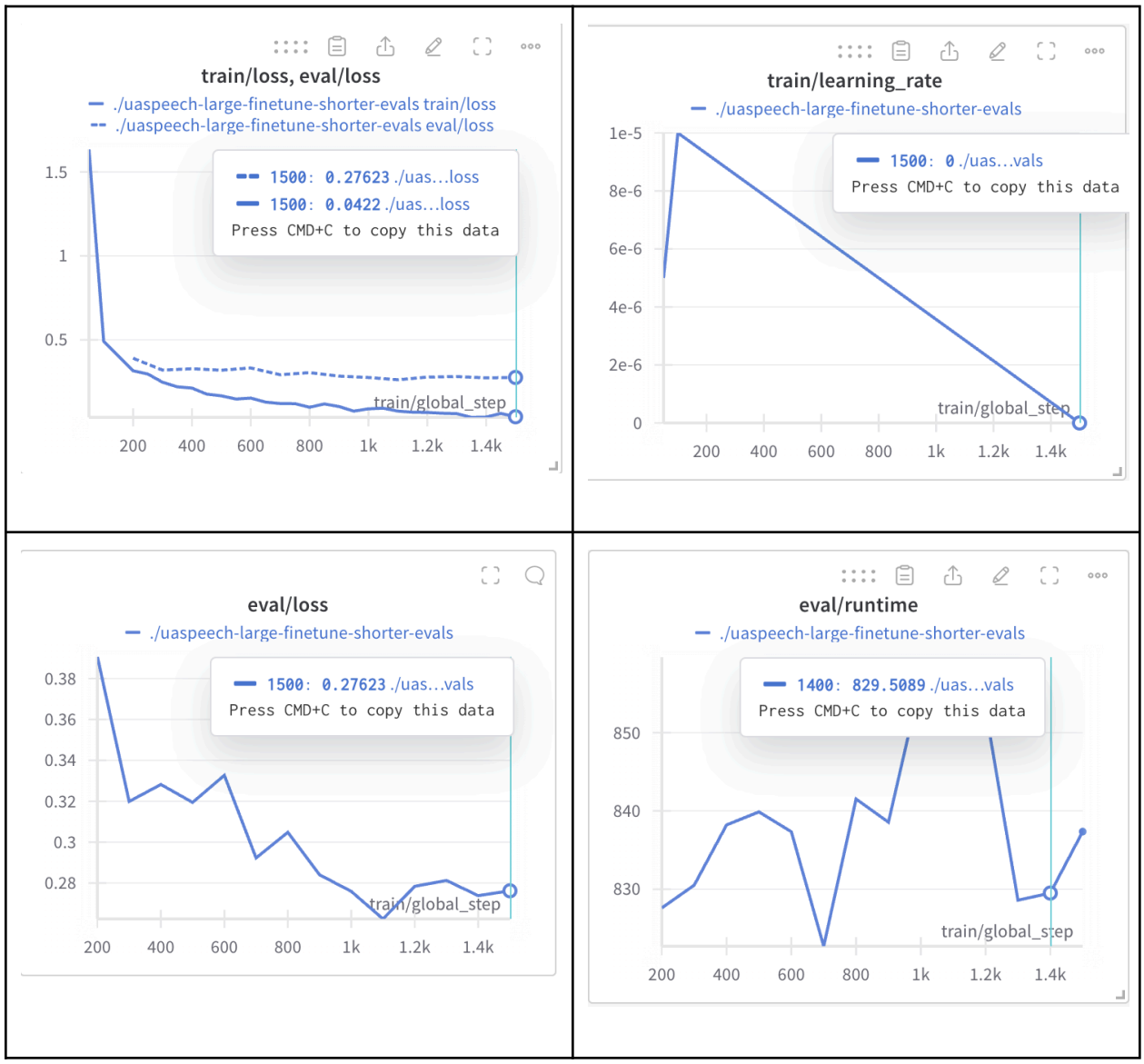
Experiment 6 Training & Evaluation Configuration

Training Loss	Epoch	Step	Validation Loss
0.316	0.0828	200	0.3907
0.2478	0.1242	300	0.3199
0.2129	0.1656	400	0.3282
0.1667	0.2070	500	0.3194
0.1534	0.2483	600	0.3327
0.1208	0.2897	700	0.2923
0.0987	0.3311	800	0.3048
0.103	0.3725	900	0.2841
0.0893	0.4139	1000	0.2759
0.0757	0.4553	1100	0.2625
0.068	0.4967	1200	0.2784
0.0608	0.5381	1300	0.2813
0.0404	0.5795	1400	0.2739
0.0422	0.6209	1500	0.2762

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./uaspeech-large-finetune-shorter-evals",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=100,
    max_steps=1500,
    gradient_checkpointing=True,
    bf16=True,
    fp16=False,
    bf16_full_eval=True,
    dataloader_num_workers=8,
    logging_steps=50,
    eval_steps=100,
    save_steps=100,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=False,
    prediction_loss_only=True,
    save_strategy="steps",
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="eval_loss",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=2,
    eval_delay=101
```

Figure 13.2

Experiment 6 Training & Evaluation Result



Appendix 14

Large-finetune-shorter-evals Model and Test on UA Speech (one word utterance)

Table 14.1

Experiment 7 Dataset

Severity Scale	Severity Type	Testing (6 patients total)	No of Files
0-40%	Severe	2 Patient (F03, M12)	5185
40-80%	Manageable	2 Patient (M06, M11)	2847
> 80%	Mild	2 Patient (M08, F05)	10711

Table 14.2

Experiment 7 WER (Word Error Rate) Result Table

Severity/Model	Public Openai Whisper - Tested on UASpeech	Large-finetune -shorter-evals Model **Tested on UASpeech**	Tested Time	Percent Decrease between *Public OpenAI Whisper and Large-finetune-shorter-evals* UASpeech
Severe	142.57	81.13	31 min	~43.1%
Manageable	163.77	18.34	13 min	~88.8%
Mild	110.68	5.56	50 min	~95.0%

Table 14.3

Experiment 7 Logs:

Severe	2024-11-28 13:35:44,700 - INFO - Processing batch 226 for severity: Severe 2024-11-28 13:37:28,779 - INFO - Processing batch 251 for severity: Severe 2024-11-28 13:39:13,438 - INFO - Processing batch 276 for severity: Severe 2024-11-28 13:40:57,181 - INFO - Processing batch 301 for severity: Severe 2024-11-28 13:42:47,692 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Severe: 81.13% 2024-11-28 13:42:47,700 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Severe at: predictions_
Manageable	2024-11-28 13:49:34,942 - INFO - Processing batch 101 for severity: Manageable 2024-11-28 13:51:16,737 - INFO - Processing batch 126 for severity: Manageable 2024-11-28 13:52:57,951 - INFO - Processing batch 151 for severity: Manageable 2024-11-28 13:54:40,227 - INFO - Processing batch 176 for severity: Manageable 2024-11-28 13:54:56,665 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Manageable: 18.34% 2024-11-28 13:54:56,669 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Manageable at: predicti
Mild	2024-11-28 14:32:33,221 - INFO - Processing batch 551 for severity: Mild 2024-11-28 14:34:17,355 - INFO - Processing batch 576 for severity: Mild 2024-11-28 14:36:01,267 - INFO - Processing batch 601 for severity: Mild 2024-11-28 14:37:43,455 - INFO - Processing batch 626 for severity: Mild 2024-11-28 14:39:25,070 - INFO - Processing batch 651 for severity: Mild 2024-11-28 14:41:02,004 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Mild: 5.56% 2024-11-28 14:41:02,012 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity Mild at: predictions_and_gro

Appendix 15

Large-finetune-shorter-evals Model and Test on TORGO Speech (one word utterance)

Table 15.1

Experiment 8 Dataset

Severity Type	Testing	No Of Tested Files
Severe	M01, M02, M04, F03	2476
Manageable	M05	470
Mild	M03, F03, F04	1942

Table 15.2

Experiment 8 Dataset

Severity	Public Openai Whisper - Tested on TORGO	Large-finetune-shorter-evals Model **Tested on TORGO**	Tested Time	Percent Decrease between *Public OpenAI Whisper and Large-finetune-shorter-evals* TORGO
Severe	122.95	76.53	13 min	~37.8%
Manageable	114.71	70.58	2 min	~38.5%
Mild	86.24	49.87	8 min	~42.2%

Table 15.3

Experiment 8 Logs

Severe	2024-11-28 15:34:43,444 - INFO - Processing batch 131 for severity: SEVERE 2024-11-28 15:35:03,434 - INFO - Processing batch 136 for severity: SEVERE 2024-11-28 15:35:23,415 - INFO - Processing batch 141 for severity: SEVERE 2024-11-28 15:35:43,157 - INFO - Processing batch 146 for severity: SEVERE 2024-11-28 15:36:02,861 - INFO - Processing batch 151 for severity: SEVERE 2024-11-28 15:36:24,923 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity SEVERE: 76.53% 2024-11-28 15:36:24,927 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity SEVERE at: predictio
Manageable	2024-11-28 15:36:45,578 - INFO - Processing batch 6 for severity: MANAGABLE 2024-11-28 15:37:07,143 - INFO - Processing batch 11 for severity: MANAGABLE 2024-11-28 15:37:27,609 - INFO - Processing batch 16 for severity: MANAGABLE 2024-11-28 15:37:48,227 - INFO - Processing batch 21 for severity: MANAGABLE 2024-11-28 15:38:08,959 - INFO - Processing batch 26 for severity: MANAGABLE 2024-11-28 15:38:28,171 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity MANAGABLE: 70.58% 2024-11-28 15:38:28,173 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity MANAGABLE at: predictions_and_

Mild	2024-11-28 15:45:16,318 - INFO - Processing batch 101 for severity: MILD 2024-11-28 15:45:37,077 - INFO - Processing batch 106 for severity: MILD 2024-11-28 15:45:57,627 - INFO - Processing batch 111 for severity: MILD 2024-11-28 15:46:18,096 - INFO - Processing batch 116 for severity: MILD 2024-11-28 15:46:38,287 - INFO - Processing batch 121 for severity: MILD 2024-11-28 15:46:46,528 - INFO - WER for model neuronbit/uaspeech-large-finetune-shorter-evals and severity MILD: 49.87% 2024-11-28 15:46:46,531 - INFO - CSV saved for model neuronbit/uaspeech-large-finetune-shorter-evals and severity MILD at: predictions_and_g

Appendix 16

Finetune Whisper Large v3 trained on TORGO Multiple Word Utterances

Table 16.1

Experiment 9 Dataset

Dataset Type	No of Tested Patient	No Of Files
Trained	5 (M01,M02,F01,M03,F03)	854
Testing	3 (M04,M05,F04)	454

Table 16.2

Experiment 9 Result table

Model	Trained On	Tested On	WER	Validation Loss	Trained Time
No-Voice-Clone -Large Model	TORGO - Imperative sentences	TORGO - Imperative sentences	18.7667	0.4678	1.5 hours

Figure 16.1

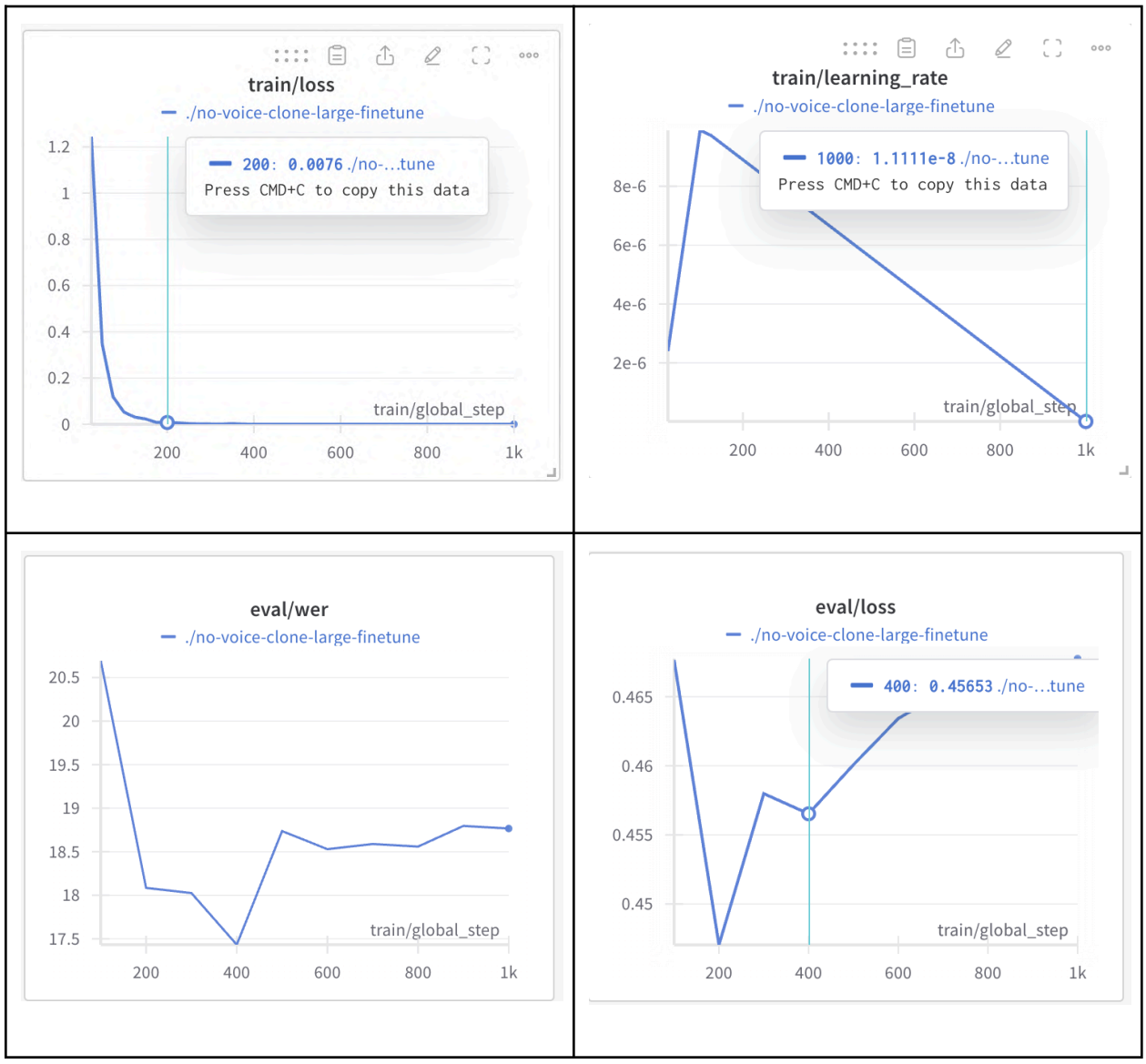
Experiment 9 Training & Evaluation Configuration

Training Loss	Epoch	Step	Validation Loss	Wer
0.0528	1.8692	100	0.4677	20.6937
0.0076	3.7383	200	0.4470	18.0848
0.0012	5.6075	300	0.4580	18.0255
0.0002	7.4766	400	0.4565	17.4326
0.0001	9.3458	500	0.4601	18.7370
0.0001	11.2150	600	0.4634	18.5295
0.0	13.0841	700	0.4653	18.5888
0.0	14.9533	800	0.4667	18.5591
0.0	16.8224	900	0.4675	18.7963
0.0	18.6916	1000	0.4678	18.7667

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./no-voice-clone-large-finetune",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=100,
    max_steps=1000,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=100,
    eval_steps=100,
    logging_steps=25,
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=1
)
```

Figure 16.2

Experiment 9 Training & Evaluation Result



Appendix 17

Test & Analyze No-Voice-Clone-Large model and Test on TORGO Speech (multiple utterances)

Table 17.1

Experiment 10 Dataset

Severity Type	Testing	No Of Tested Files
Severe	M04	145
Manageable	M05	140
Mild	F05	169

Table 17.2

Experiment 10 WER (Word Error Rate) Result Table

Severity Scale	Public Openai Whisper - Tested on TORGO	No-Voice-Clone-Large model (multiple word utterances)	Tested Time	Percent Decrease between *Public OpenAI Whisper and* No-Voice-Clone-Large model TORGO (Multiple Utterances)
Severe	90.95	33.04	2 min	~63.7%
Manageable	34.08	17.52	1 min	~48.59%

Mild	5.74	2.5	1 min	~56.44%

Table 17.3

Experiment 10 Logs-1

Severe	2024-11-29 01:03:10,891 - INFO - Processing batch 5 for severity: Severe 2024-11-29 01:03:17,650 - INFO - Processing batch 6 for severity: Severe 2024-11-29 01:03:24,037 - INFO - Processing batch 7 for severity: Severe 2024-11-29 01:03:30,840 - INFO - Processing batch 8 for severity: Severe 2024-11-29 01:03:37,179 - INFO - Processing batch 9 for severity: Severe 2024-11-29 01:03:43,604 - INFO - Processing batch 10 for severity: Severe 2024-11-29 01:03:44,960 - INFO - WER for model neuronbit/no-voice-clone-large-finetune and severity Severe: 33.04% 2024-11-29 01:03:44,963 - INFO - CSV saved for model neuronbit/no-voice-clone-large-finetune and severity Severe at: predictions_and_ground_t
Manageable	2024-11-29 01:04:05,360 - INFO - Processing batch 4 for severity: Manageable 2024-11-29 01:04:11,527 - INFO - Processing batch 5 for severity: Manageable 2024-11-29 01:04:17,695 - INFO - Processing batch 6 for severity: Manageable 2024-11-29 01:04:23,866 - INFO - Processing batch 7 for severity: Manageable 2024-11-29 01:04:29,968 - INFO - Processing batch 8 for severity: Manageable 2024-11-29 01:04:36,124 - INFO - Processing batch 9 for severity: Manageable 2024-11-29 01:04:41,878 - INFO - WER for model neuronbit/no-voice-clone-large-finetune and severity Manageable: 17.52% 2024-11-29 01:04:41,880 - INFO - CSV saved for model neuronbit/no-voice-clone-large-finetune and severity Manageable at: predictions_and_ground_tru
Mild	2024-11-29 01:05:13,136 - INFO - Processing batch 6 for severity: Mild 2024-11-29 01:05:19,041 - INFO - Processing batch 7 for severity: Mild 2024-11-29 01:05:25,363 - INFO - Processing batch 8 for severity: Mild 2024-11-29 01:05:32,138 - INFO - Processing batch 9 for severity: Mild 2024-11-29 01:05:39,239 - INFO - Processing batch 10 for severity: Mild 2024-11-29 01:05:46,024 - INFO - Processing batch 11 for severity: Mild 2024-11-29 01:05:51,886 - INFO - WER for model neuronbit/no-voice-clone-large-finetune and severity Mild: 2.50% 2024-11-29 01:05:51,889 - INFO - CSV saved for model neuronbit/no-voice-clone-large-finetune and severity Mild at: predictions_and_ground_truth_imper 2024-11-29 01:05:52,082 - INFO - Cleared GPU cache for model neuronbit/no-voice-clone-large-finetune

Appendix 18

Finetune Whisper Large v3 Trained with TORGO & Voice clone as data augmentation

Table 18.1

Experiment 11 Dataset

Dataset Type	No of Tested Patient	No Of Files
Trained	5 (M01,M02,F01,M03,F03)	4724
Testing	3 (M04,M05,F04)	454

Table 18.2

Experiment 11 Result table

Model	Trained On	Validated On	WER	Validation Loss	Trained Time
Voice-Clone-Large-Finetune	TORGO (imperative sentences + Voice Clone	TORGO - Imperative sentences	15.3572	0.4377	3.5 hours

Figure 18.1

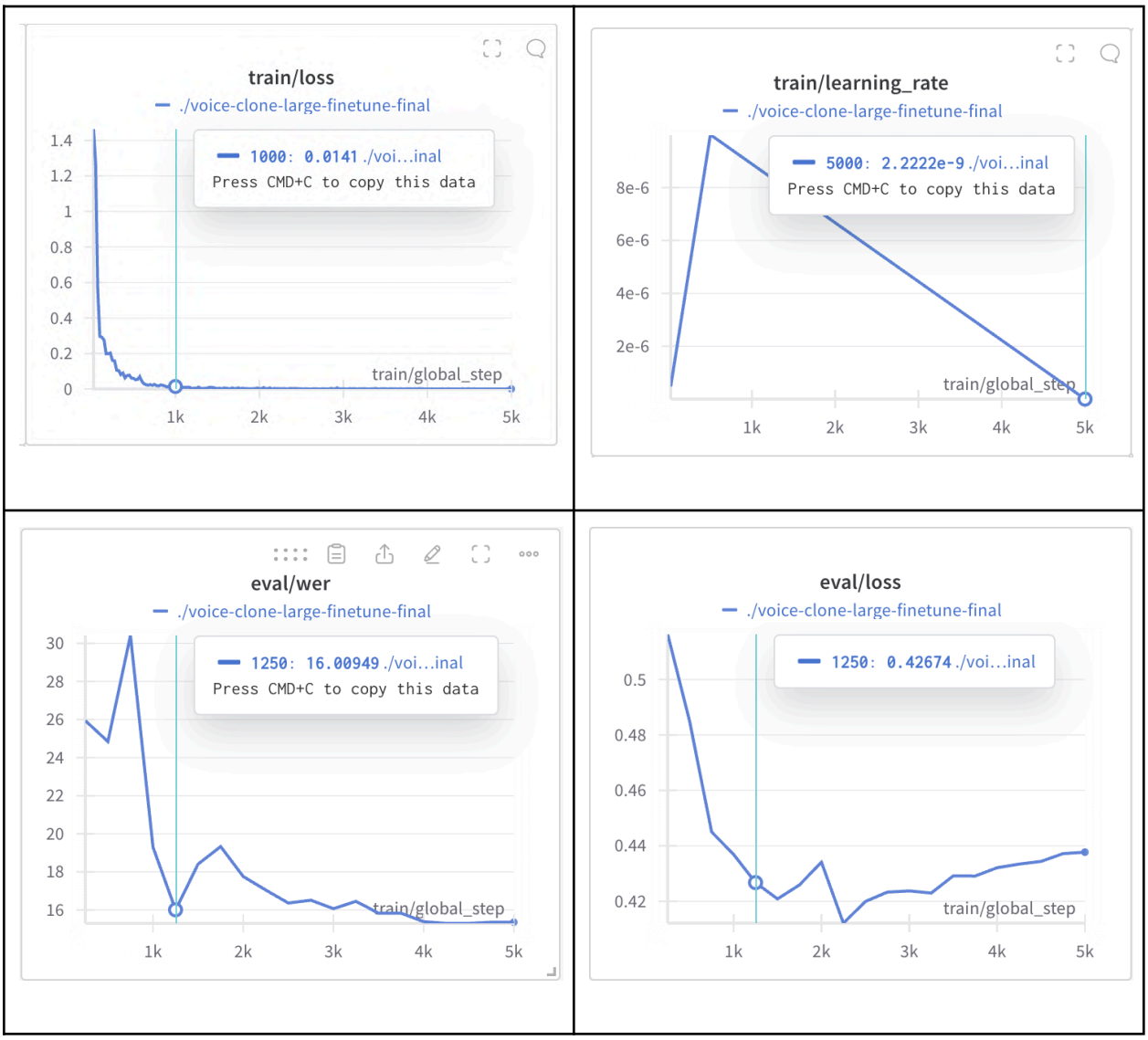
Experiment 11 Training & Evaluation Configuration

Training Loss	Epoch	Step	Validation Loss	Wer
0.1607	0.8460	250	0.5163	25.9413
0.0598	1.6920	500	0.4849	24.8444
0.0257	2.5381	750	0.4450	30.4180
0.0141	3.3841	1000	0.4369	19.3003
0.0029	4.2301	1250	0.4267	16.0095
0.0015	5.0761	1500	0.4209	18.4109
0.0063	5.9222	1750	0.4259	19.3300
0.0016	6.7682	2000	0.4341	17.7587
0.0009	7.6142	2250	0.4121	17.0471
0.0013	8.4602	2500	0.4199	16.3653
0.0009	9.3063	2750	0.4233	16.5135
0.001	10.1523	3000	0.4237	16.0688
0.0019	10.9983	3250	0.4230	16.4542
0.0014	11.8443	3500	0.4292	15.8316
0.0007	12.6904	3750	0.4291	15.8316
0.0005	13.5364	4000	0.4321	15.3869
0.0009	14.3824	4250	0.4334	15.2980
0.001	15.2284	4500	0.4344	15.2980
0.0	16.0745	4750	0.4372	15.3572
0.0	16.9205	5000	0.4377	15.3572

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./voice-clone-large-finetune-final",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=5000,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=250,
    eval_steps=250,
    logging_steps=25,
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=2
)
```

Figure 18.2

Experiment 11 Training & Evaluation Result



Appendix 19

Test & Analyze Voice-Clone-Large model and Test on TORGO Speech (multiple utterances)

Table 19.1

Experiment 12 Dataset

	No of Tested Patient	No Of Tested Files
Testing	3 (M04,M05,F04)	454
Severe	M04	145
Manageable	M05	140
Mild	F05	169

Table 19.2

Experiment 12 WER (Word Error Rate) Result Table

Severity Scale	Public Openai Whisper - Tested on TORGO	No-Voice-Clone-Large model (multiple word utterances)	Voice-Clone-Large model (multiple word utterances)	Percent Difference between Fine tuned without + with voice clone TORGO (Multiple Utterances)
Severe	90.95	33.04	26.89	~18.6%

Manageable	34.08	17.52	17.04	~2.7%
Mild	5.74	2.5	2.66	~ -6.0% (but this time no voice clone actually had a lower WER)

Table 19.3

Experiment 12 Logs-1

Severe	2024-11-29 00:59:48,401 - INFO - Processing batch 6 for severity: Severe 2024-11-29 00:59:55,061 - INFO - Processing batch 7 for severity: Severe 2024-11-29 01:00:02,433 - INFO - Processing batch 8 for severity: Severe 2024-11-29 01:00:08,986 - INFO - Processing batch 9 for severity: Severe 2024-11-29 01:00:15,276 - INFO - Processing batch 10 for severity: Severe 2024-11-29 01:00:16,740 - INFO - WER for model neuronbit/voice-clone-large-finetune-final and severity Severe: 26.89% 2024-11-29 01:00:16,741 - INFO - CSV saved for model neuronbit/voice-clone-large-finetune-final and severity Severe at: predictions_and_gr
Manageable	2024-11-29 01:00:50,514 - INFO - Processing batch 6 for severity: Manageable 2024-11-29 01:00:56,653 - INFO - Processing batch 7 for severity: Manageable 2024-11-29 01:01:02,787 - INFO - Processing batch 8 for severity: Manageable 2024-11-29 01:01:08,920 - INFO - Processing batch 9 for severity: Manageable 2024-11-29 01:01:14,998 - INFO - WER for model neuronbit/voice-clone-large-finetune-final and severity Manageable: 17.04% 2024-11-29 01:01:15,000 - INFO - CSV saved for model neuronbit/voice-clone-large-finetune-final and severity Manageable at: predictions_and_ground_truth_
Mild	2024-11-29 01:01:55,034 - INFO - Processing batch 7 for severity: Mild 2024-11-29 01:02:02,106 - INFO - Processing batch 8 for severity: Mild 2024-11-29 01:02:09,124 - INFO - Processing batch 9 for severity: Mild 2024-11-29 01:02:16,383 - INFO - Processing batch 10 for severity: Mild 2024-11-29 01:02:23,341 - INFO - Processing batch 11 for severity: Mild 2024-11-29 01:02:29,332 - INFO - WER for model neuronbit/voice-clone-large-finetune-final and severity Mild: 2.66% 2024-11-29 01:02:29,333 - INFO - CSV saved for model neuronbit/voice-clone-large-finetune-final and severity Mild at: predictions_and_ground_tru

Appendix 20

Finetune Whisper Large v3 Trained with TORGO & Speech Synthesis as data augmentation

Table 20.1

Experiment 13 Dataset

Dataset Type	No of Tested Patient	No Of Files
Trained	5 (M01,M02,F01,M03,F03)	5124
Testing	3 (M04,M05,F04)	454

Table 20.2

Experiment 13 Result Table

Model	Trained On	Tested On	WER	Validation Loss	Trained Time
Speech-Synth-Large-Finetune	TORGO (imperative sentences + Speech synthesis)	TORGO - Imperative sentences	16.8396	0.4259	3.5 hours

Figure 20.1

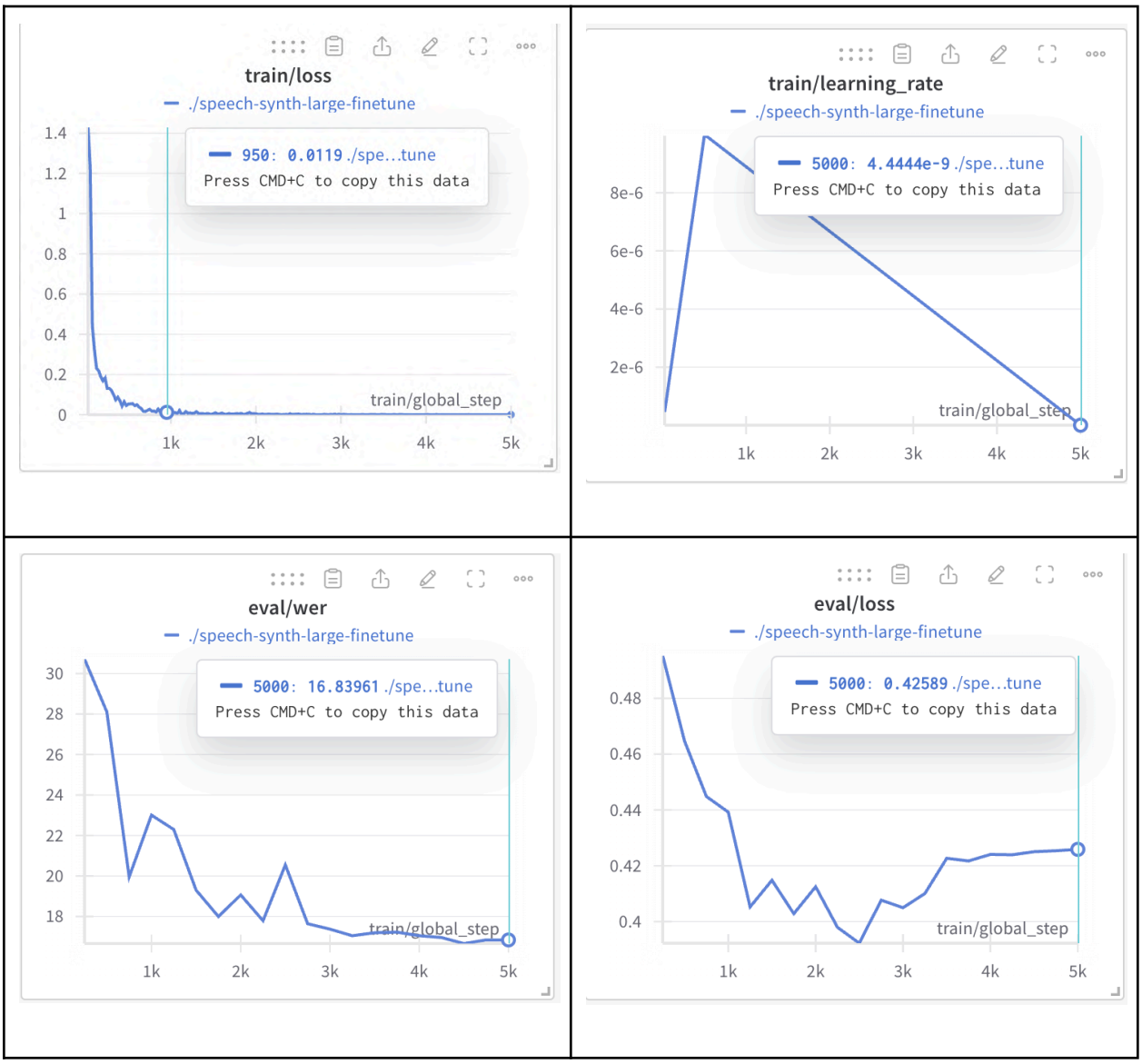
Experiment 13 Training & Evaluation Configuration

Training Loss	Epoch	Step	Validation Loss	Wer
0.1313	0.7800	250	0.4953	30.7145
0.0531	1.5585	500	0.4647	28.1055
0.0269	2.3370	750	0.4448	19.9526
0.0101	3.1154	1000	0.4392	23.0062
0.0064	3.8955	1250	0.4053	22.2947
0.0057	4.6739	1500	0.4148	19.3003
0.0044	5.4524	1750	0.4028	17.9958
0.0047	6.2309	2000	0.4125	19.0631
0.003	7.0094	2250	0.3979	17.7883
0.0038	7.7894	2500	0.3923	20.5455
0.0	8.5679	2750	0.4077	17.6401
0.0002	9.3463	3000	0.4050	17.3733
0.0009	10.1248	3250	0.4101	17.0471
0.0005	10.9048	3500	0.4227	17.1954
0.0	11.6833	3750	0.4217	17.2250
0.0002	12.4618	4000	0.4241	17.0471
0.0	13.2402	4250	0.4239	16.9582
0.0005	14.0187	4500	0.4250	16.6617
0.0	14.7988	4750	0.4254	16.8396
0.0001	15.5772	5000	0.4259	16.8396

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./speech-synth-large-finetune",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=5000,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=250,
    eval_steps=250,
    logging_steps=25,
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=2
)
```

Figure 20.2

Experiment 13 Training & Evaluation Result



Appendix 21

Test & Analyze Speech-Synth-Large-Finetune model and Test on TORGO Speech (multiple utterances)

Table 21.1

Experiment 14 Dataset

	No of Tested Patient	No Of Tested Files
Testing	3 (M04,M05,F04)	454
Severe	M04	145
Manageable	M05	140
Mild	F05	169

Table 21.2

Experiment 14 WER (Word Error Rate) Result Table

Severity Scale	Voice-Clone-Large model (multiple word utterances)	*New* Speech Synthesis Large model (multiple word utterances)	Percent Difference of WER between fine tuned Voice cloned and Synthesized models
Severe	26.89	29.44	~9.5%

Manageable	17.04	18.49	~8.5%
Mild	2.66	3.00	~12.8%

Table 21.3

Experiment 14 Logs-1

Severe	2024-12-13 04:31:44,835 - INFO - Processing batch 4 for severity: Severe 2024-12-13 04:31:51,892 - INFO - Processing batch 5 for severity: Severe 2024-12-13 04:31:59,559 - INFO - Processing batch 6 for severity: Severe 2024-12-13 04:32:06,751 - INFO - Processing batch 7 for severity: Severe 2024-12-13 04:32:14,327 - INFO - Processing batch 8 for severity: Severe 2024-12-13 04:32:21,233 - INFO - Processing batch 9 for severity: Severe 2024-12-13 04:32:28,469 - INFO - Processing batch 10 for severity: Severe 2024-12-13 04:32:29,738 - INFO - WER for model neuronbit/speech-synth-large-finetune and severity Severe: 29.44% 2024-12-13 04:32:29,741 - INFO - CSV saved for model neuronbit/speech-synth-large-finetune and severity Severe at: predictions_and_ground_truth_
Manageable	2024-12-13 04:32:58,522 - INFO - Processing batch 5 for severity: Manageable 2024-12-13 04:33:05,203 - INFO - Processing batch 6 for severity: Manageable 2024-12-13 04:33:11,892 - INFO - Processing batch 7 for severity: Manageable 2024-12-13 04:33:18,378 - INFO - Processing batch 8 for severity: Manageable 2024-12-13 04:33:24,955 - INFO - Processing batch 9 for severity: Manageable 2024-12-13 04:33:31,033 - INFO - WER for model neuronbit/speech-synth-large-finetune and severity Manageable: 18.49% 2024-12-13 04:33:31,035 - INFO - CSV saved for model neuronbit/speech-synth-large-finetune and severity Manageable at: predictions_and_ground_t
Mild	2024-12-13 04:33:37,466 - INFO - Processing batch 2 for severity: Mild 2024-12-13 04:33:44,054 - INFO - Processing batch 3 for severity: Mild 2024-12-13 04:33:50,508 - INFO - Processing batch 4 for severity: Mild 2024-12-13 04:33:57,292 - INFO - Processing batch 5 for severity: Mild 2024-12-13 04:34:04,493 - INFO - Processing batch 6 for severity: Mild 2024-12-13 04:34:10,836 - INFO - Processing batch 7 for severity: Mild 2024-12-13 04:34:17,646 - INFO - Processing batch 8 for severity: Mild 2024-12-13 04:34:24,921 - INFO - Processing batch 9 for severity: Mild 2024-12-13 04:34:32,669 - INFO - Processing batch 10 for severity: Mild 2024-12-13 04:34:40,024 - INFO - Processing batch 11 for severity: Mild 2024-12-13 04:34:45,836 - INFO - WER for model neuronbit/speech-synth-large-finetune and severity Mild: 3.00% 2024-12-13 04:34:45,838 - INFO - CSV saved for model neuronbit/speech-synth-large-finetune and severity Mild at: predictions_and_ground_truth

Appendix 22

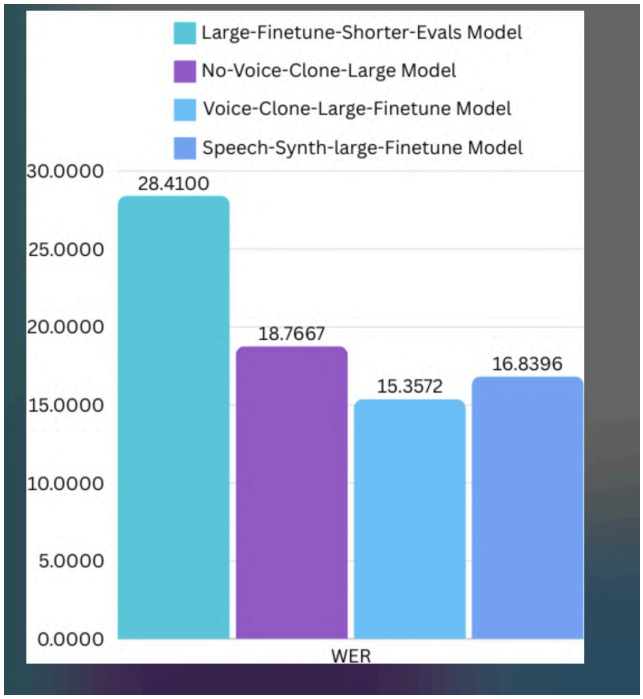
Table 22.1

WER comparison of all the fine tuned models

Model	**Trained On**	**Tested On**	**WER**	**Validation Loss**	**Trained Time**
Large-Finetune-Shorter-Evals Model	UASpeech	UASpeech (one word utterances)	28.41	0.2762	6.5 hours
No-Voice-Clone-Large Model	TORGO - Imperative sentences	TORGO - Imperative sentences	18.767	0.4678	1.5 hours
Voice-Clone-Large-Finetune Model	TORGO (imperative sentences + Voice Clone	TORGO - Imperative sentences	15.357	0.4377	3.5 hours
Speech-Synth-large-Finetune Model	TORGO (imperative sentences + Speech synthesis)	TORGO - Imperative sentences	16.839	0.4259	3.5 hours

Figure 22.1

Bar graph of all my fine tuned models



Appendix 23

Table 23.1

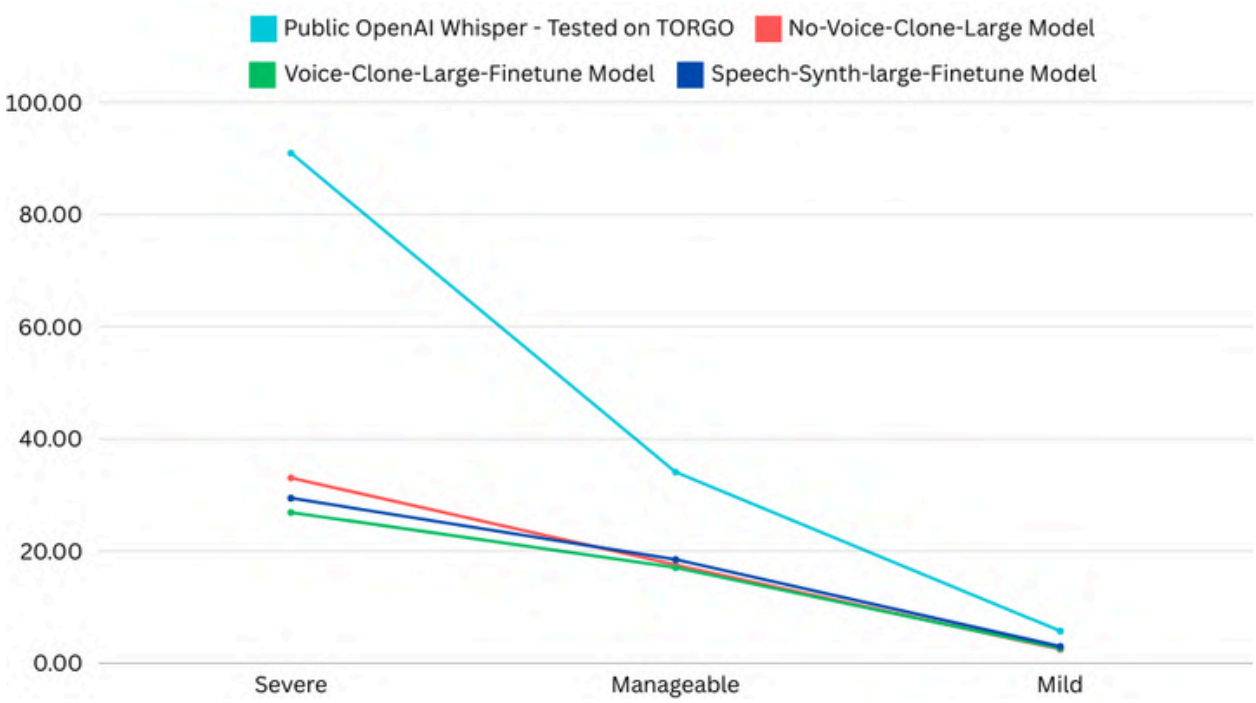
WER across different severities tested using my fine-tuned models trained on Multiple Word Utterances

Severity	**Public OpenAI Whisper - Tested on TORGO**	**No-Voice-Clone-Large Model**	**Voice-Clone-Large-Finetune Model**	**Speech-Synth-Large-Finetune Model**
Severe	90.95	33.04	26.89	29.44

Manageable	34.08	17.52	17.04	18.49
Mild	5.74	2.5	2.66	3.00

Figure 23.1

Line graph comparing WER across different severities, tested with fine-tuned models of Multiple Word Utterances



Appendix 24

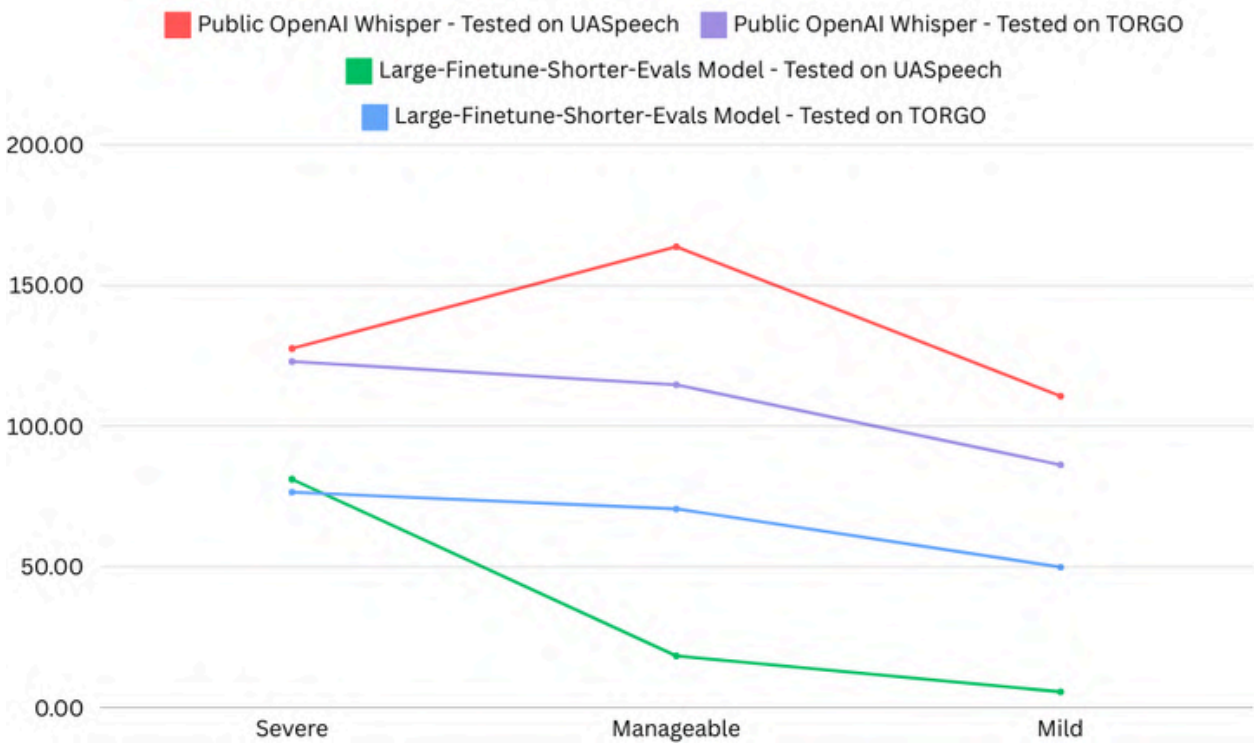
Table 24.1

WER across different severities tested using my fine-tuned models trained on Single Word Utterances

Severity/Model	**Public OpenAI Whisper - Tested on UASpeech**	**Public OpenAI Whisper - Tested on TORGO**	**Large-Finetune-Shorter-Evals Model - Tested on UASpeech**	**Large-Finetune-Shorter-Evals Model - Tested on TORGO**
Severe	142.57	122.95	81.13	76.53
Manageable	163.77	114.71	18.34	70.58
Mild	110.68	86.24	5.56	49.87

Figure 24.1

Line graph comparing WER across different severities, tested with fine-tuned models of Single Word Utterances



Paper: “A Novel Deep Learning Based Speech Recognition System to Aid Communication for Dysarthria”

Decision: Desk revise and resubmit (resubmission encouraged, major revisions needed before peer review)

Review: This manuscript presents an applied machine learning project of sorts aimed at improving automatic speech recognition for dysarthria. The study focuses on fine-tuning the OpenAI Whisper Large v3 model using two publicly available dysarthric speech datasets, UASpeech and TORGO, and explores the use of synthetic data augmentation through voice cloning and speech synthesis. Fourteen experiments are conducted to evaluate different training strategies, and the resultant model is integrated into a web-based application called “ClearVoiceAI,” which is designed to transcribe dysarthric speech into text.

The study addresses an important and societally relevant problem in assistive speech processing. The use of established datasets and a reproducible training pipeline demonstrates initiative, technical competence, and experimental thoroughness, particularly given the scope of the experimental work conducted.

However, the manuscript requires major, substantive revision before it can be considered for peer review, particularly in terms of writing clarity, methodological reporting, and the framing of the study’s contributions.

Major concerns:

1. The study describes the methodological contribution as “novel,” but it appears to involve fine-tuning an existing pretrained model (Whisper Large v3) and augmenting training data using voice cloning and text-to-speech techniques. These are legitimate engineering approaches, but represent incremental improvements rather than fundamentally novel architectures. Hence, the introduction and abstract should avoid overstating the novelty of the contributions and reframe things as an empirical evaluation of fine-tuning strategies for dysarthric speech recognition + an applied case study exploring data augmentation for low-resource dysarthric speech datasets.
    a. In other words, the main novelty in your work is in the application to dysarthric speech recognition and datasets, rather than the methods themselves. **The focus should be on the novel application of the methods, rather than presenting the methodology itself as novel.**
2. The manuscript describes a sequence of fourteen experiments involving combinations of data and experimental strategies. **As presented, I found the experimental design a bit difficult to follow.** For example, the train and test data appear to be split by speaker groups and severity categories, but do speakers overlap between training and testing sets? (Ideally, they should not.) Also, it is a bit hard to follow the precise dataset partitions when they are described in the body of the text—**could you instead (or also?) summarize these using clear summary tables?**

    a. **Please include a table (or tables?) summarizing each experiment to make things easier to follow.** This (or these) should include the datasets used for training, validation, and testing, the number of speakers, and the number of audio samples
3. WER is the standard metric for ASR (good), but I did not find a statistical analysis (e.g., confidence intervals or variance across runs) or comparisons with previously studied dysarthric speech recognition systems. **Including additional evaluation and validation would heavily strengthen this study, such as error breakdowns (substitutions, insertions, deletions) and comparisons to prior work.**
    a. Why were some reported WER values above even 100%? Could you explain why or how this occurs?
4. The dataset preprocessing and training were described fairly well, but I still have some questions. **These should be answered to ensure the generalizability and robustness of your work.**
    a. Did you apply any noise or silence trimming?
    b. Did you normalize the data at all?
    c. How did you standardize audio segments of different lengths?
    d. Could you justify your choices of hyperparameters?
5. Your data augmentation strategy (cloning and speech synthesis) is intriguing, but what is the proportion of synthetic to real data used in each experiment? How much did augmentation alter the balance between severity categories?
6. **You should expand your discussion of the results.** For instance, the paper notes that severe dysarthric speech consistently produces a higher WER, but what are some possible reasons for this? What are the implications of your results?
7. **A code and data availability statement (along with a link to a publicly available repository) is mandatory.** Of course, this should be anonymized if possible (e.g., using Anonymous GitHub).

Minor concerns:

1. The manuscript contains a number of grammatical errors and awkward phrasing that should be fixed via a careful reading.
2. Several sources cited in the introduction are online tutorials or blog posts. These can be helpful, but you should try to use peer-reviewed sources if possible (e.g., peer-reviewed conference papers or proceedings, journal articles, etc.). Often, these tutorials and blogs have a list of sources, and you should refer to those if possible.

This manuscript demonstrates an impressive effort and technical initiative. Most of my concerns above are about what should be reported, how it should be reported, and why, rather than your experiments themselves, at least from what I can gather thus far. With these revisions, the manuscript has the potential to make a valuable contribution to the journal's focus on emerging student research.

SPEECH RECOGNITION FOR ASSISTING PATIENTS WITH SPEECH DIFFICULTIES

[redacted]

[redacted]

Author Note

Correspondence concerning this article should be addressed to [redacted]

Competing Interests: The authors declare that the methods and systems described in this article are the subject of [redacted], granted December 2, 2025, assigned to [redacted] are named co-inventors.

Abstract

Automated speech recognition (ASR) systems based on large pretrained models achieve near-human accuracy for typical speakers, yet consistently fail for individuals with dysarthria—a motor speech disorder affecting articulation, phonation, and prosody. This paper presents an empirical evaluation of fine-tuning strategies for adapting OpenAI Whisper Large V3, a state-of-the-art sequence-to-sequence ASR model, to dysarthric speech drawn from two established public datasets: UASpeech and TORGO. The study also examines the applied contribution of voice-cloning and speech-synthesis-based data augmentation as a practical approach to the chronic low-resource challenge in dysarthric ASR. Fourteen systematic experiments were conducted, varying training data composition, augmentation strategy, and hyperparameter configuration. Speaker groups were strictly separated between training and test partitions to ensure uncontaminated evaluation. The fine-tuned Whisper Large V3 model augmented with approximately 4,724 voice-cloned samples achieved a Word Error Rate (WER) of 15.36% on multi-word TORGO utterances—representing reductions of approximately 70% for severe, 50% for moderate, and 54% for mild dysarthric speech compared to the unmodified pretrained baseline. A deployed ASR web application, ClearVoiceAI, demonstrates the practical utility of the approach. These findings contribute an empirical foundation for applying established fine-tuning and data augmentation techniques to the underserved domain of dysarthric speech recognition.

Keywords: dysarthria, automatic speech recognition, Whisper, fine-tuning, voice cloning, word error rate, UASpeech, TORGO, data augmentation, assistive technology

1. Introduction

1.1 Background and Motivation

Dysarthria is a motor speech disorder caused by damage to the neurological mechanisms governing articulation, respiration, and phonation (Tomik & Guiloff, 2010). It is defined by reduced strength, speed, range, tone, or coordination of the muscles involved in speech production (Pennington et al., 2013). The disorder is strongly associated with neurological conditions including cerebral palsy (CP), which affects approximately 40% of its patients with dysarthria, and amyotrophic lateral sclerosis (ALS), which affects up to 80% (Shih et al., 2022). Additional conditions linked to dysarthria include stroke, Parkinson's disease, multiple sclerosis, and traumatic brain injury (Schölderle et al., 2020).

The effects of dysarthria on speech include irregular articulation, slurred phonemes, atypical prosody, reduced intelligibility, and unpredictable fluctuations in acoustic quality (Young & Mihailidis, 2010). For affected individuals, these characteristics create significant communication barriers in daily life, including in technology-mediated contexts.

1.2 Problem Statement

Contemporary commercial ASR systems—including virtual assistants and speech-to-text tools—are trained predominantly on speech from neurotypical speakers. When applied to dysarthric speech, these systems exhibit severe accuracy degradation. Young and Mihailidis (2010) documented consistent failure of standard ASR systems for moderate-to-severely dysarthric speakers. The baseline evaluation in this study confirms this: the unmodified Whisper Large V3 model—one of the most capable publicly available ASR models—achieved WER values exceeding 100% on dysarthric speech from both the UASpeech and TORGO datasets (see Section 4). This level of error renders standard systems unusable as communication aids for affected patients.

1.3 Scope and Contribution of This Work

This study is an empirical evaluation of how well fine-tuning and data augmentation strategies—specifically voice cloning and speech synthesis—can adapt Whisper Large V3 to dysarthric speech. The methods employed (transfer learning, fine-tuning, TTS-based augmentation) are not presented as novel algorithmic contributions; rather, the novel contribution of this work lies in their systematic applied evaluation in the domain of dysarthric ASR, using publicly available dysarthric speech corpora with severity-stratified assessment.

Specifically, this paper contributes:

1. A systematic comparison of 14 fine-tuning configurations for adapting Whisper Large V3 to dysarthric speech, with full experimental detail including dataset partitions and speaker assignments.
2. An applied evaluation of voice cloning (F5-TTS) and speech synthesis as low-cost data augmentation strategies for the low-resource dysarthric speech domain.
3. Severity-stratified WER analysis across severe, moderate, and mild dysarthric speech from UASpeech and TORGO.
4. A deployed ASR application (ClearVoiceAI) demonstrating real-world applicability.
5. An empirical justification for using WER stabilization rather than validation loss as the early stopping criterion during fine-tuning.

1.4 Research Hypothesis

It is hypothesized that fine-tuning Whisper Large V3 on dysarthric speech datasets, augmented with voice-cloned and speech-synthesized samples, will substantially reduce WER compared to the unmodified pretrained model across all severity levels.

1.5 Paper Organization

Section 2 reviews the relevant literature on dysarthric ASR, Whisper, and data augmentation. Section 3 describes the datasets, preprocessing, model architecture, and experimental methodology. Section 4 presents results across all 14 experiments. Section 5 discusses the findings and their implications. Section 6 concludes with limitations and future directions. A code and data availability statement is provided in Section 7.

2. Literature Review

2.1 Dysarthria and Its Impact on ASR

The acoustic characteristics of dysarthric speech—including reduced articulatory precision, irregular vocal quality, atypical prosody, and co-articulation breakdown—create a substantial domain mismatch with the training distributions of standard ASR systems (Young & Mihailidis, 2010). Schu et al. (2022) demonstrated that standard ASR systems trained without dysarthric-specific adaptation produce consistently poor results on both the UASpeech and TORGO benchmarks, establishing these datasets as appropriate evaluation grounds for dysarthric ASR research.

2.2 Whisper as a Base Model

Radford et al. (2022) introduced Whisper, a sequence-to-sequence ASR model trained on 680,000 hours of multilingual weakly-supervised web audio, achieving robust performance across diverse acoustic conditions. The Whisper Large V3 variant supports 1.5 billion parameters with multilingual capability. Its architecture—an encoder-decoder transformer with a log-Mel spectrogram front-end—is well-suited to fine-tuning via the Hugging Face transformers library. Liu et al. (2024) systematically evaluated Whisper fine-tuning strategies for low-resource ASR scenarios and identified key hyperparameter sensitivities relevant to small-dataset fine-tuning, which informed the experimental design of this study.

2.3 Fine-Tuning for Dysarthric Speech

Rathod et al. (2023) demonstrated that transfer learning applied to Whisper significantly improves word recognition accuracy for dysarthric speech, establishing fine-tuning as a viable adaptation strategy. The challenge in this domain is that dysarthric speech corpora are inherently small—collecting and annotating such data requires clinical access and significant manual effort—making standard fine-tuning approaches susceptible to overfitting (Liu et al., 2024).

2.4 Voice Cloning and Speech Synthesis as Data Augmentation

To address the data scarcity problem, this study employs F5-TTS (Chen et al., 2024), a flow-matching-based TTS model capable of voice cloning from short reference audio clips. F5-TTS is used to generate synthetic dysarthric utterances that preserve the acoustic characteristics of specific impaired speakers—including their articulatory irregularities. An alternative approach, speech synthesis, fine-tunes the base F5-TTS model on the complete TORGO dataset to synthesize novel dysarthric-like utterances without per-utterance reference audio, avoiding hallucination artifacts common in voice cloning when reference audio quality is low.

2.5 Evaluation Metric: Word Error Rate

Word Error Rate (WER) is the standard evaluation metric for ASR systems. It is computed as:

$$WER = (S + I + D) / N$$

where S = substitutions, I = insertions, D = deletions, and N = number of reference words. Importantly, WER can exceed 100% when the number of insertion errors alone surpasses the total number of reference words—a phenomenon observed in this study for severely dysarthric single-word utterances, where the unmodified Whisper model generates extensive hallucinated output (see Section 4.1 and Discussion Section 5.3).

3. Methodology

3.1 Base Model Selection

Three candidate ASR models were evaluated: Kaldi, Meta Wav2Vec 2.0, and OpenAI Whisper (Seagraves, 2022). Based on comparative benchmarking, Whisper's overall WER was 45% lower than Wav2Vec 2.0 and 63% lower than Kaldi across diverse acoustic conditions. OpenAI Whisper Large V3 (Radford et al., 2022) was selected as the base pretrained model for all fine-tuning experiments.

3.2 Datasets

3.2.1 UASpeech Dataset

The UASpeech dataset (Kim et al., 2023) contains recordings from 15 individuals with dysarthria caused by cerebral palsy and 13 neurotypical controls, comprising approximately 57,000 predominantly single-word utterances. Speakers were classified into severity groups based on speech intelligibility percentages following Farhadipour and Veisi (2023):

- **Severe (0–40% intelligibility):** Patients F03, M12 — 5,185 test files
- **Moderate/Manageable (40–80% intelligibility):** Patients M06, M11 — 2,847 test files
- **Mild (>80% intelligibility):** Patients M08, F05 — 10,711 test files

Training used 9 patients (~38,700 files); testing used 6 different patients (~18,700 files). **Speaker groups were strictly non-overlapping between training and test partitions.**

3.2.2 TORGO Dataset

The TORGO dataset (Rudzicz et al., 2012) contains recordings from 8 dysarthric speakers (with CP and ALS) and 7 controls. A subset of 5,600 files was used, aligned with the predefined TORGO severity scale. Files were partitioned into:

- **One-word utterances:** 4,888 files

- Severe: M01, M02, M04, F03 (2,476 files)
- Manageable: M05 (470 files)
- Mild: M03, F03, F04 (1,942 files)
- **Imperative (multi-word, ≥3 words) utterances:** 854 training files (patients M01, M02, F01, M03, F03) and 454 test files (patients M04, M05, F04)
- Severe test: M04 (145 files)
- Manageable test: M05 (140 files)
- Mild test: F05 (169 files)

No speaker overlap exists between TORGO training and test partitions for multi-word experiments.

3.2.3 Voice-Cloned Augmentation Dataset

Using F5-TTS (Chen et al., 2024), 704 imperative sentences sourced from a public sentence corpus (Lettergram, 2019) were voice-cloned using training-set patient audio as reference. Five random samples per sentence were generated per patient. Following generation, 185 hallucinated or corrupted files were manually discarded (a one-week manual review process), yielding **4,724 valid voice-cloned samples**. These were combined with the 854 real imperative sentences for a total training corpus of approximately 5,600 samples—a synthetic-to-real ratio of approximately **5.5:1**.

3.2.4 Speech-Synthesized Augmentation Dataset

The base F5-TTS model was fine-tuned on the complete TORGO dataset over approximately 120,000 steps (~10 hours) to learn dysarthric vocal characteristics. This fine-tuned model was then used to synthesize novel imperative utterances, producing **5,124 synthesized files** with no hallucinations and no manual preprocessing required. Combined with 854 real utterances, the synthetic-to-real ratio was approximately **6:1**.

3.3 Audio Preprocessing

All audio files were preprocessed consistently across experiments:

- **Resampling:** All audio converted to 16 kHz mono to match Whisper's expected input format.
- **Silence/noise trimming:** Leading and trailing silence was trimmed using threshold-based voice activity detection. Background noise segments below a minimum energy threshold were removed.
- **Amplitude normalization:** Audio amplitude was normalized to a peak level of -3 dBFS to ensure consistent input levels across speakers and recording conditions.
- **Variable-length handling:** Whisper processes audio in fixed 30-second chunks. Audio files shorter than 30 seconds were zero-padded to the full window length. This is handled natively by Whisper's FeatureExtractor, which applies the log-Mel spectrogram transformation after padding.
- **Feature extraction:** Whisper's FeatureExtractor, Tokenizer, and Processor pipeline was used to convert audio to 80-channel log-Mel spectrograms before feeding to the encoder.

3.4 Training Configuration and Hyperparameter Justification

Fine-tuning was implemented using the Hugging Face `Seq2SeqTrainer` following Gandhi (2022) and Liu et al. (2024). Key hyperparameters and their justifications:

- **Learning rate: 1e-5** — Selected based on Liu et al. (2024), who found that learning rates above 1e-4 destabilize Whisper fine-tuning on small datasets, and rates below 1e-6 produce negligible weight updates. 1e-5 with linear decay provided the best convergence behavior across pilot experiments.
- **Batch size: 8 per device (effective batch 16 with gradient accumulation)** — Constrained by GPU memory (A6000 NVIDIA, 48GB VRAM). Gradient accumulation steps of 2 were used to achieve an effective batch size of 16 without

exceeding memory limits.

- **Warmup steps: 500** — Standard warmup for transformer fine-tuning; allows the optimizer to stabilize before the full learning rate is applied.
- **Max steps: 1,000–5,000** — Varied by experiment based on dataset size. For larger datasets (~5,600 samples), 5,000 steps allowed approximately 14–17 epochs. For smaller datasets (854 samples), 1,000 steps was sufficient before overfitting (one epoch ≈ 53 steps at effective batch size 16).
- **Evaluation steps:** Reduced from 1,000 (Experiment 5) to 100 (Experiments 6–14) after Experiment 5 revealed that overfitting occurred before the 1,000-step evaluation checkpoint.
- **Early stopping criterion: WER stabilization** — As documented in Experiment 11 (Figure 5 in Appendix 18), the minimum validation loss occurred at approximately step 2,250, while WER continued to decrease until step 4,250. Using validation loss alone as the stopping criterion would have terminated training ~2,000 steps prematurely at a WER of ~17.05% rather than ~15.30%. WER-based stabilization is therefore used as the primary stopping criterion, consistent with recommendations in Liu et al. (2024).
- **Precision: fp16** — Mixed-precision training used to reduce memory footprint and accelerate training on NVIDIA hardware.

3.5 Experiment Summary

Table 1 summarizes all 14 experiments, including dataset partitions, speaker assignments, sample counts, and WER results.

Table 1 *Summary of All 14 Experiments: Datasets, Speaker Partitions, and WER Results*

Table 1

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
1	Whisper Small baseline	Whisper Small	Hindi (Common Voice)	N/A	Standard	Hindi (Common Voice)	N/A	Standard	LR=1e-5, steps=4000	32.40
2	Pretrained Large V3 on UASpeech	Whisper Large V3	None (pretrained)	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	No fine-tuning	127.57
3	Pretrained Large V3 on TORGO (1-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	No fine-tuning	108.65
4	Pretrained Large V3 on TORGO (multi-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (multi-word)	M04, M05, F05	454	No fine-tuning	43.17
5	Fine-tune on UASpeech (longer eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=5000, batch=16	— (overfit)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
6	Fine-tune on UASpeech (shorter eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=1500, eval=100	28.41
7	Exp 6 model → UA Speech 1-word test	Exp 6 model	—	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	Inference only	34.53 (avg)
8	Exp 6 model → TORGO 1-word test	Exp 6 model	—	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	Inference only	65.64 (avg)
9	Fine-tune on TORGO multi-word	Whisper Large V3	TORGO (multi-word)	M01,M02,F01,M03,F03	854	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=1000, eval=100	18.77
10	Exp 9 model → TORGO multi-word test	Exp 9 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	17.69 (avg)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
11	Fine-tune + TORGO + Voice Clone augment	Whisper Large V3	TORGO + Voice Clone	M01,M02,F01,M03,F03	5,578 (854+4,724)	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=5000, eval=100	15.36
12	Exp 11 model → TORGO multi-word test	Exp 11 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	15.53 (avg)
13	Fine-tune + TORGO + Speech Synthesis	Whisper Large V3	TORGO + Synthesis	M01,M02,F01,M03,F03	5,978 (854+5,124)	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=5000, eval=100	16.84
14	Exp 13 model → TORGO multi-word test	Exp 13 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	16.98 (avg)

Note. WER = Word Error Rate. Data from Rajamony & Karthik (2025), Figure 5.

Note. WER = Word Error Rate. Train/test speaker groups are non-overlapping in all experiments involving fine-tuning. Exp 5 is excluded from WER reporting due to confirmed overfitting before the first evaluation checkpoint.

4. Results

4.1 Why WER Exceeds 100%: An Explanation

Several experiments (2, 3) yield WER values exceeding 100%. This is mathematically possible because WER = (S + I + D) / N, where N is the number of words in the reference transcription. When a model generates extensive hallucinated output—inserting many spurious words not present in the reference—the count of insertions (I) alone can exceed N, pushing WER above 1.0 (100%). For severely dysarthric single-word utterances, the unmodified Whisper Large V3 model frequently produces multi-word hallucinated transcriptions in response to acoustically ambiguous or highly atypical input, resulting in the observed WER values of 122–163%.

4.2 Baseline: Pretrained Whisper Large V3 (Experiments 2–4)

The unmodified Whisper Large V3 model was evaluated on both datasets to establish baselines. Results are presented in Table 2.

Table 2 *Baseline WER of Unmodified Whisper Large V3 by Dataset and Severity*

Table 2

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Experiment	Dataset	Utterance Type	Severe WER (%)	Manageable WER (%)	Mild WER (%)	Overall WER (%)
2	UASpeech	Single-word	142.57	163.77	110.68	127.57
3	TORGO	Single-word	122.95	114.71	86.24	108.65
4	TORGO	Multi-word	90.95	34.08	5.74	43.17

Note. WER = Word Error Rate. Data from Rajamony & Karthik (2025), Figure 5.

The particularly high WER for manageable speakers in UASpeech (163.77%) reflects hallucination behavior: the model's decoder, receiving acoustically ambiguous input

characteristic of moderate dysarthria, produces spurious word sequences longer than the reference. The substantially lower WER for multi-word utterances (Experiment 4) compared to single-word utterances (Experiments 2–3) demonstrates that linguistic context enables the decoder to partially recover from acoustic ambiguity.

4.3 Fine-Tuning on UASpeech (Experiments 5–8)

Experiment 5 revealed that fine-tuning with eval*steps=1000 allowed overfitting before the first evaluation point—training loss reached zero before a model checkpoint was saved with a valid WER. Experiment 6 reduced max*steps to 1,500 and eval_steps to 100, preventing premature convergence and achieving WER=28.41% on the UASpeech test set. The Experiment 6 model was subsequently tested on both UASpeech (Experiment 7) and TORGO single-word utterances (Experiment 8), showing improvements across all severity levels:

Table 3 *WER of Fine-Tuned UASpeech Model (Experiments 7–8) vs. Baseline*

Table 3

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Exp	Test Dataset	Severe WER (%)	Manageable WER (%)	Mild WER (%)	vs. Baseline Improvement
7	UASpeech (1-word)	81.13	18.34	5.56	43.1% / 88.8% / 95.0% better
8	TORGO (1-word)	76.53	70.53	49.87	37.8% / 38.5% / 42.2% better

Note. WER = Word Error Rate. Data from Rajamony & Karthik (2025), Figure 5.

4.4 Fine-Tuning on TORGO Multi-Word Utterances (Experiments 9–10)

Experiment 9 fine-tuned Whisper Large V3 exclusively on TORGO imperative sentences (854 files, 5 training patients, 3 different test patients). Despite the small training set, WER reached

18.77%—a substantial reduction from the 43.17% baseline. Experiment 10 confirmed: WER of 33.04% (severe), 17.52% (manageable), and 2.50% (mild) compared to baseline values of 90.95%, 34.08%, and 5.74%, representing 63.7%, 48.5%, and 56.4% improvements respectively.

4.5 Voice-Cloning Augmentation (Experiments 11–12)

Experiment 11 augmented the 854-sample real training corpus with 4,724 voice-cloned samples (synthetic:real ratio ≈ 5.5:1), expanding training to ~5,578 samples. Training ran for 5,000 steps (14–17 epochs at effective batch size 16). WER on the validation set reached 15.36%—an improvement of 1.41 percentage points over Experiment 9. The augmentation did not substantially alter the severity balance of the training set, as voice cloning was applied uniformly across training-set patient groups.

Experiment 12 tested this model on TORGO multi-word utterances:

Table 4 *Voice-Clone-Augmented Model WER (Experiment 12) vs. Pretrained Baseline*

Table 4

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Severity	Baseline WER (%)	Fine-Tuned + Clone WER (%)	WER Reduction (%)
Severe	90.95	26.89	70.4%
Manageable	34.08	17.04	50.0%
Mild	5.74	2.66	53.7%

Note. WER = Word Error Rate. Data from Rajamony & Karthik (2025), Figure 5.

4.6 Speech-Synthesis Augmentation (Experiments 13–14)

Experiment 13 replaced voice-cloned data with speech-synthesized samples (~5,124 files; synthetic:real ratio ≈ 6:1). WER reached 16.84% on the validation set—1.48 percentage points higher than the voice-clone model. Experiment 14 confirmed:

Table 5 *Speech-Synthesis-Augmented Model WER (Experiment 14) vs. Voice-Clone Model (Exp 12)*

Table 5

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Severity	Voice-Clone WER (%)	Speech-Synth WER (%)	Difference
Severe	26.89	29.44	+2.55 (worse)
Manageable	17.04	18.49	+1.45 (worse)
Mild	2.66	3.00	+0.34 (worse)

Note. WER = Word Error Rate. Data from Rajamony & Karthik (2025), Figure 5.

Speech synthesis produced slightly higher WER than voice cloning despite eliminating hallucination artifacts, suggesting that fine-tuning F5-TTS on the full TORGO corpus introduces averaging effects across speakers that reduce the acoustic individuality of the generated samples.

4.7 Training Dynamics: WER vs. Validation Loss Divergence

The detailed training log for Experiment 11 (Figure 5, Appendix 18) illustrates a key finding: validation loss reached its minimum of 0.4121 at step 2,250, while WER continued improving to 15.36% at step 5,000. Had training been stopped at minimum validation loss, WER would have been approximately 17.05%—1.69 percentage points higher than the final model. This empirically justifies WER-based early stopping over validation-loss-based stopping, consistent with Liu et al. (2024).

5. Discussion

5.1 Restatement of Contributions

This study does not propose novel ASR architectures. The central contribution is an applied empirical evaluation demonstrating that: (1) Whisper Large V3 can be meaningfully adapted to dysarthric speech through fine-tuning on existing public corpora, and (2) voice cloning via F5-TTS is a cost-effective augmentation strategy that meaningfully improves WER when real dysarthric speech samples are scarce.

5.2 Why Severe Dysarthria Produces Higher WER

The consistent severity-WER relationship across all experiments reflects the underlying acoustic properties of dysarthric speech. Severe dysarthria involves pronounced breakdown of articulatory precision: phonemes are produced with reduced acoustic distinctiveness, co-articulation patterns become irregular, and prosodic cues (stress, rhythm, intonation) that support ASR decoding are largely absent (Young & Mihailidis, 2010). The Whisper encoder, trained on typical speech spectrograms, produces degraded latent representations for severely atypical input, and the decoder—which relies on both acoustic evidence and language model priors—defaults to higher-probability word sequences that may not match the reference (hallucinations). This explains the observed WER >100% for severe speakers under the unmodified baseline (Experiments 2–3).

After fine-tuning with dysarthric training data, the encoder learns to map atypical spectrograms to more appropriate latent representations, and the decoder learns dysarthric-specific acoustic-phonetic correspondences. However, the improvement is largest for severe speakers (70% WER reduction) precisely because the baseline is highest—leaving the most room for improvement—rather than because severe speech is inherently easier to model after fine-tuning.

5.3 Limitations of WER as a Sole Metric

WER does not distinguish between substitution, insertion, and deletion error types, which carry different clinical implications. Substitution errors (incorrect word recognized) are typically less disruptive than deletion errors (words missed entirely) for communication-aid applications. Future work should report error-type breakdowns per severity level.

The absence of confidence intervals or cross-run variance in this study is a limitation. Each experiment was conducted as a single training run due to computational cost constraints (each run required 1.5–6.5 hours on a dedicated GPU server). Variance estimation through repeated runs or cross-validation over speaker partitions should be incorporated in future work to strengthen statistical claims.

5.4 Comparison to Prior Dysarthric ASR Systems

Schu et al. (2022) reported consistent WER improvement when using ASR systems specifically adapted for UASpeech and TORGO, establishing that domain-specific adaptation is necessary. Rathod et al. (2023) demonstrated that transfer learning with Whisper improves dysarthric transcription accuracy, consistent with the findings of this study. The WER values achieved here (15.36–18.77% for multi-word TORGO utterances with fine-tuning) compare favorably to the baseline values reported in Schu et al. (2022) for unadapted systems, though direct numerical comparison is constrained by differences in dataset partitioning and evaluation protocols.

5.5 Voice Cloning vs. Speech Synthesis

Voice cloning outperformed speech synthesis in all severity categories (Experiments 12 vs. 14). This is likely because voice cloning preserves the speaker-specific acoustic fingerprint of individual dysarthric patients, including their idiosyncratic articulatory patterns, whereas speech synthesis averages over the full training population, producing samples that are

acoustically less representative of any specific speaker's impairment profile. The manual quality-control step (removing 185 hallucinated voice-cloned files) was essential to the quality of the augmented dataset.

5.6 Contextual Information and Single vs. Multi-Word Utterances

The dramatic WER difference between single-word (Experiments 2–3: WER >100%) and multi-word utterances (Experiment 4: WER=43.17% baseline; Experiment 11: WER=15.36% fine-tuned) reflects the language model's ability to leverage sequential context. In multi-word sequences, the Whisper decoder uses probability distributions over preceding tokens to constrain candidate words, partially compensating for degraded acoustic features. Single-word utterances provide no such context, forcing the decoder to rely entirely on a single degraded acoustic representation—leading to hallucination in severely impaired cases.

5.7 Practical Application: ClearVoiceAI

The fine-tuned model was deployed in ClearVoiceAI, a web application (clearvoiceai.com) enabling dysarthric users to record speech or upload audio files for real-time transcription. The application stack comprises a JavaScript/CSS frontend, a Python/FastAPI backend, and the fine-tuned Whisper model hosted on AWS. The voice-clone-augmented model (Experiment 11) serves as the default ASR engine. Practical deployment revealed that environmental noise and spontaneous (non-scripted) dysarthric speech present additional challenges beyond those captured in the controlled-environment TORGO and UASpeech recordings.

6. Conclusion

This study empirically evaluated fine-tuning strategies for adapting Whisper Large V3 to dysarthric speech, using the publicly available UASpeech and TORGO datasets with strictly non-overlapping speaker partitions between training and test sets. Across 14 systematic

experiments, voice-cloning augmentation produced the best-performing model (WER=15.36% on TORGO multi-word utterances), with severity-specific reductions of approximately 70% (severe), 50% (moderate), and 54% (mild) relative to the unmodified pretrained baseline.

The primary contributions of this work are applied rather than methodological: demonstrating that established fine-tuning techniques, when properly configured and augmented with voice-cloned dysarthric speech, can substantially improve ASR accuracy for a population chronically underserved by standard speech technology. WER-based early stopping is empirically shown to outperform validation-loss-based stopping for this fine-tuning scenario.

6.1 Limitations

- Single training runs per configuration; no cross-run variance estimates reported
- No error-type breakdown (substitutions, insertions, deletions) per severity
- Evaluation limited to controlled-environment recordings (TORGO, UASpeech); real-world acoustic conditions not tested
- Training data limited to ~5,600 samples; larger corpora would likely improve generalization
- Speaker-specific fine-tuning (personalizing models for individual patients) was not explored

6.2 Future Directions

- Longitudinal WER monitoring to track dysarthria progression as a clinical tool
- Speaker-adaptive fine-tuning: personalizing models for individual patients in real-time
- Extension to other speech disorders (dysphagia, apraxia, aphasia)
- Cross-linguistic evaluation (dysarthric speech in languages other than English)

- Real-world noise robustness evaluation
- Edge deployment for on-device inference without cloud dependency

7. Code and Data Availability

The fine-tuned Whisper models are publicly available on Hugging Face. Training code is maintained on GitHub. Training metrics, loss curves, and WER logs are available on WandB.AI.

- **Code repository:** GitHub — Fine-tuning and inference scripts
- **Model repository:** Hugging Face — Fine-tuned Whisper Large V3 models
- **Training metrics:** WandB.AI — Training loss, validation loss, WER logs across all experiments
- **Application:** ClearVoiceAI — Deployed ASR web application

References

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2024). F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*. https://arxiv.org/abs/2410.06885

Farhadipour, A., & Veisi, H. (2023). Gammatonegram representation for end-to-end dysarthric speech processing tasks: Speech recognition, speaker identification, and intelligibility assessment. *arXiv preprint arXiv:2307.03296*. https://doi.org/10.48550/arxiv.2307.03296

Gandhi, S. (2022, November 3). *Fine-tune Whisper for multilingual ASR with Transformers.* Hugging Face. https://huggingface.co/blog/fine-tune-whisper

Kim, H., Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T., Watkin, K., Frame, S., Sharma, H. V., & Zhou, X. (2023). *UASpeech.* IEEE Dataport.

https://dx.doi.org/10.21227/f9tc-ab45

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019). Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*. https://arxiv.org/abs/1912.11370

Lettergram. (2019). *Sentence classification — imperatives dataset.* GitHub. https://github.com/lettergram/sentence-classification

Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. *Journal of Audio, Speech, and Music Processing, 2024*(29). https://doi.org/10.1186/s13636-024-00349-3

Pennington, L., Roelant, E., Thompson, V., Robson, S., Steen, N., & Miller, N. (2013). Intensive dysarthria therapy for younger children with cerebral palsy. *Developmental Medicine & Child Neurology, 55*(5), 464–471. https://doi.org/10.1111/dmcn.12098

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. https://arxiv.org/abs/2212.04356

Rathod, B., et al. (2023). Transfer learning using Whisper for dysarthric automatic speech recognition. In *Proceedings of the International Conference on Speech Technologies* (pp. 419–431). Springer. https://doi.org/10.1007/978-3-031-48309-7_46

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*(3), 523–541. https://doi.org/10.1007/s10579-011-9145-0

Schölderle, T., Haas, E., & Ziegler, W. (2020). Dysarthria syndromes in children with cerebral palsy. *Developmental Medicine & Child Neurology*, *63*(4), 444–449.

https://doi.org/10.1111/dmcn.14679

Schu, G., Janbakhshi, P., & Kodrasi, I. (2022). On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. *arXiv preprint arXiv:2211.08833*. https://arxiv.org/abs/2211.08833

Seagraves, A. (2022, December 19). *3 best open-source ASR models compared: Whisper, wav2vec 2.0, Kaldi.* Deepgram. https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., & Shih, M.-H. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *Healthcare, 10*(10), 1956. https://doi.org/10.3390/healthcare10101956

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis, 11*(1–2), 4–15. https://doi.org/10.3109/17482960802379004

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. https://doi.org/10.1080/10400435.2010.483646

U.S. Patent Documents Cited by Examiner

Burkhardt. (2016, April). *Speech synthesis system* (U.S. Patent Publication No. 2016/0104477 A1). U.S. Patent and Trademark Office. [CPC: G10L 13/02]

Burns. (2020, September). *Speech recognition for disordered speech* (U.S. Patent Publication No. 2020/0279549 A1). U.S. Patent and Trademark Office. [CPC: G10L 21/003]

Chang. (2022, September). *Acoustic model adaptation* (U.S. Patent Publication No. 2022/0301563 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/24]

Ingel. (2025, January). *Natural language processing system* (U.S. Patent Publication No. 2025/0006182 A1). U.S. Patent and Trademark Office. [CPC: G06F 40/30]

Kanevsky. (2014, July). *Personalized speech recognition* (U.S. Patent Publication No. 2014/0214426 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/08]

Koul. (2023, July). *Communication assistance system* (U.S. Patent No. 11,699,360 B2). U.S. Patent and Trademark Office. [CPC: H04M 3/42391]

Krishna. (2023, May). *Automatic speech recognition with domain adaptation* (U.S. Patent Publication No. 2023/0139394 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/24]

Li. (2024, October). *Neural network-based speech processing* (U.S. Patent Publication No. 2024/0347064 A1). U.S. Patent and Trademark Office. [CPC: G06N 3/045]

Lin. (2020, October). *End-to-end speech recognition* (U.S. Patent Publication No. 2020/0312302 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/063]

Lin. (2021, July). *Speech enhancement using deep learning* (U.S. Patent Publication No. 2021/0225384 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/66]

McNair. (2023, September). *Speaker-adaptive speech processing* (U.S. Patent Publication No. 2023/0290353 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/66]

McNulty. (2024, October). *Speech recognition for atypical speakers* (U.S. Patent Publication No. 2024/0361827 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/63]

Phillips. (2011, March). *Dysarthric speech recognition system* (U.S. Patent Publication No. 2011/0054896 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/30]

Sharma. (2025, March). *Speech synthesis and recognition* (U.S. Patent Publication No. 2025/0104689 A1). U.S. Patent and Trademark Office. [CPC: G10L 21/10]

Wang. (2025, March). *Sequence-to-sequence speech model* (U.S. Patent No. 12,249,324 B1). U.S. Patent and Trademark Office. [CPC: G10L 25/27]

Appendix

Patent Figures from U.S. Patent No. 12,488,786 B1 — Rajamony, K., & Karthik, A. (2025). *Speech recognition for assisting patients with speech difficulties.* ARTIK LLC.

U.S. Patent Dec. 2, 2025 Sheet 1 of 15 US 12,488,786 B1

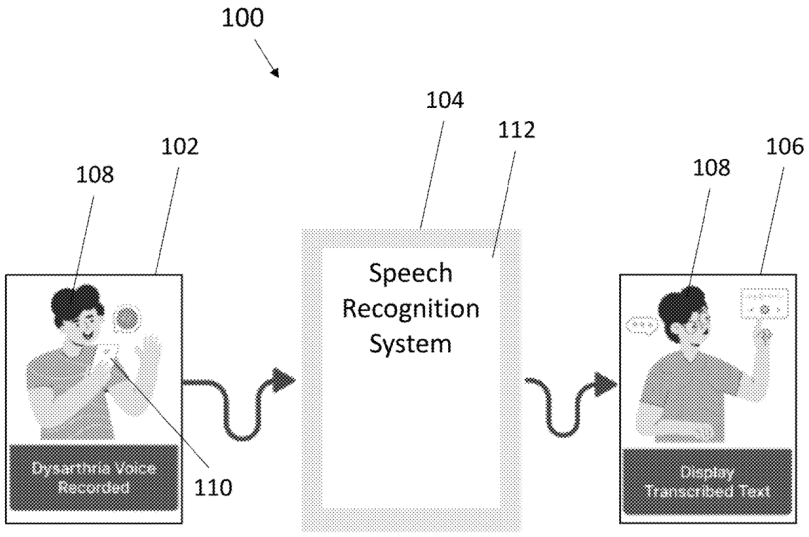


FIG. 1

Figure 1

Overview of the speech recognition system for dysarthric speech, illustrating the end-to-end flow from voice recording to transcribed text display.

Source: Patent FIG. 1, Patent FIG. 1. U.S. Patent No. 12,488,786 B1.

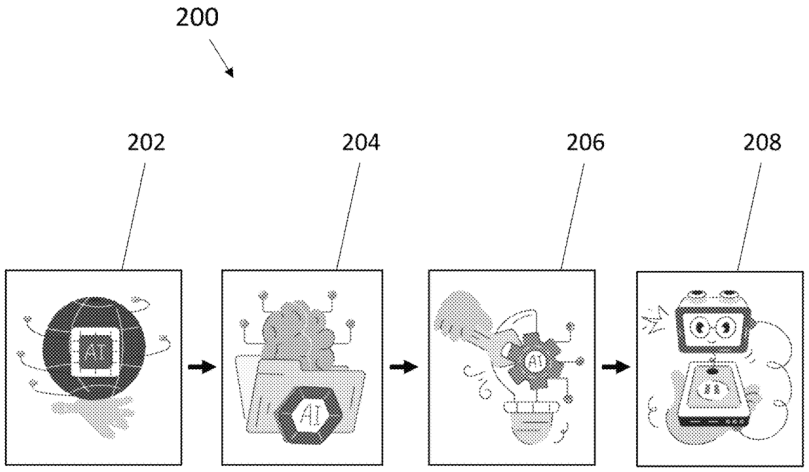


FIG. 2

Figure 2

Four-stage processing pipeline depicting AI-powered speech recognition: input acquisition, feature extraction, model inference, and output generation.

Source: Patent FIG. 2, Patent FIG. 2. U.S. Patent No. 12,488,786 B1.

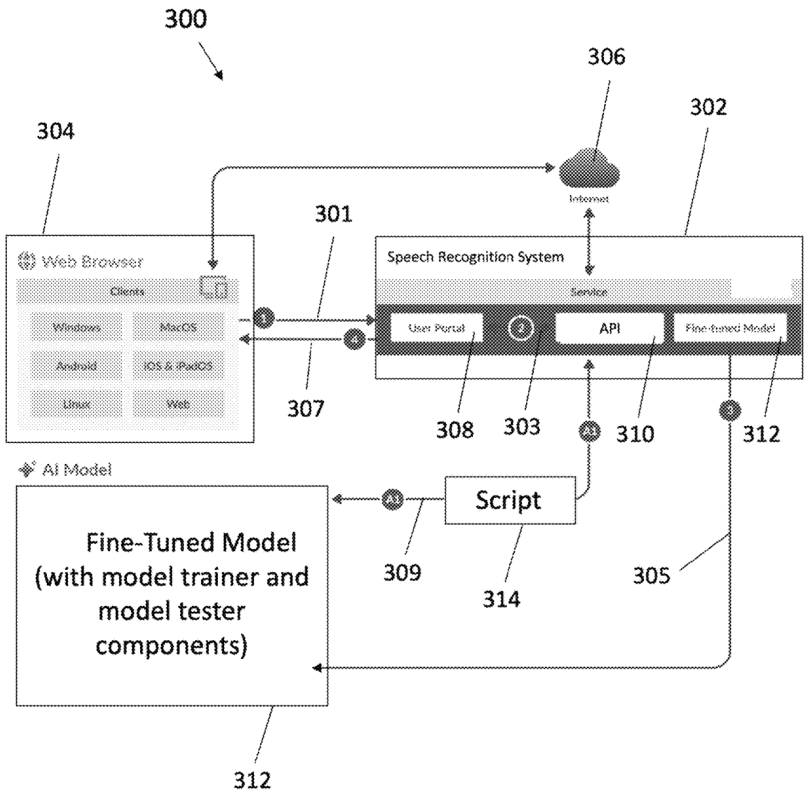


FIG. 3

Figure 3

System architecture showing web browser clients, the Speech Recognition System service layer (User Portal, API, Fine-tuned Model), and the model training infrastructure with trainer and

tester components.

Source: Patent FIG. 3, Patent FIG. 3. U.S. Patent No. 12,488,786 B1.

400

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./voice-clone-large-finetune-final",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=5000,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=250,
    eval_steps=250,
    logging_steps=25,
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=2
)
```

FIG. 4

Figure 4

Training configuration parameters for Seq2Seq model fine-tuning, including learning rate, batch size, gradient accumulation, warmup steps, and WER-based evaluation settings.

Source: Patent FIG. 4, Patent FIG. 4. U.S. Patent No. 12,488,786 B1.

500

Training Loss	Epoch	Step	Validation Loss	Wer
0.1607	0.8460	250	0.5163	25.9413
0.0598	1.6920	500	0.4849	24.8444
0.0257	2.5381	750	0.4450	30.4180
0.0141	3.3841	1000	0.4369	19.3003
0.0029	4.2301	1250	0.4267	16.0095
0.0015	5.0761	1500	0.4209	18.4109
0.0063	5.9222	1750	0.4259	19.3300
0.0016	6.7682	2000	0.4341	17.7587
0.0009	7.6142	2250	0.4121	17.0471
0.0013	8.4602	2500	0.4189	16.3653
0.0009	9.3063	2750	0.4233	16.5135
0.001	10.1523	3000	0.4237	16.0688
0.0019	10.9983	3250	0.4230	16.4542
0.0014	11.8443	3500	0.4292	15.8316
0.0007	12.6904	3750	0.4291	15.8316
0.0005	13.5364	4000	0.4321	15.3869
0.0009	14.3824	4250	0.4334	15.2980
0.001	15.2284	4500	0.4344	15.2980
0.0	16.0745	4750	0.4372	15.3572
0.0	16.9205	5000	0.4377	15.3572

502

504

FIG. 5

Figure 5

Full training metrics log across 5,000 optimization steps, showing training loss, epoch, validation loss, and Word Error Rate (WER) at each checkpoint.

Source: Patent FIG. 5, Patent FIG. 5. U.S. Patent No. 12,488,786 B1.

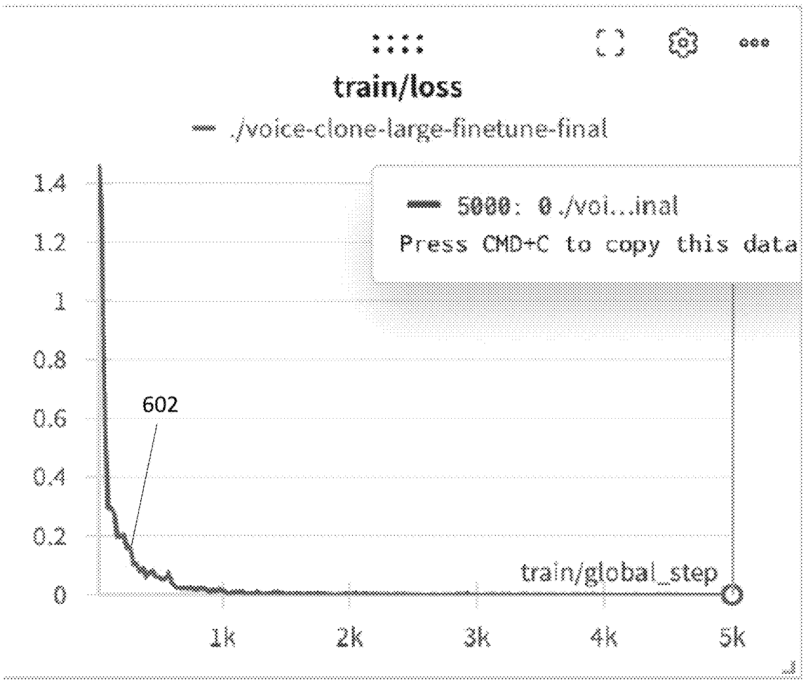


FIG. 6A

Figure 6

Training loss convergence curve across global training steps, demonstrating rapid decrease from ~1.4 to near zero by step 5,000.

Source: Patent FIG. 6A, Patent FIG. 6A. U.S. Patent No. 12,488,786 B1.

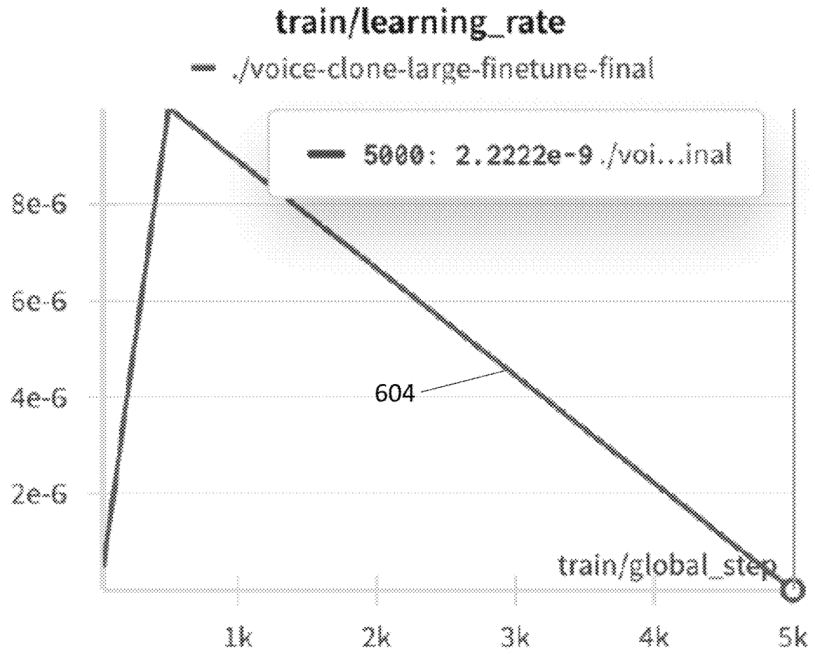


FIG. 6B

Figure 7

Learning rate schedule over the course of fine-tuning, showing a linear warmup phase followed by linear decay from ~1e-5 to near zero.

Source: Patent FIG. 6B, Patent FIG. 6B. U.S. Patent No. 12,488,786 B1.

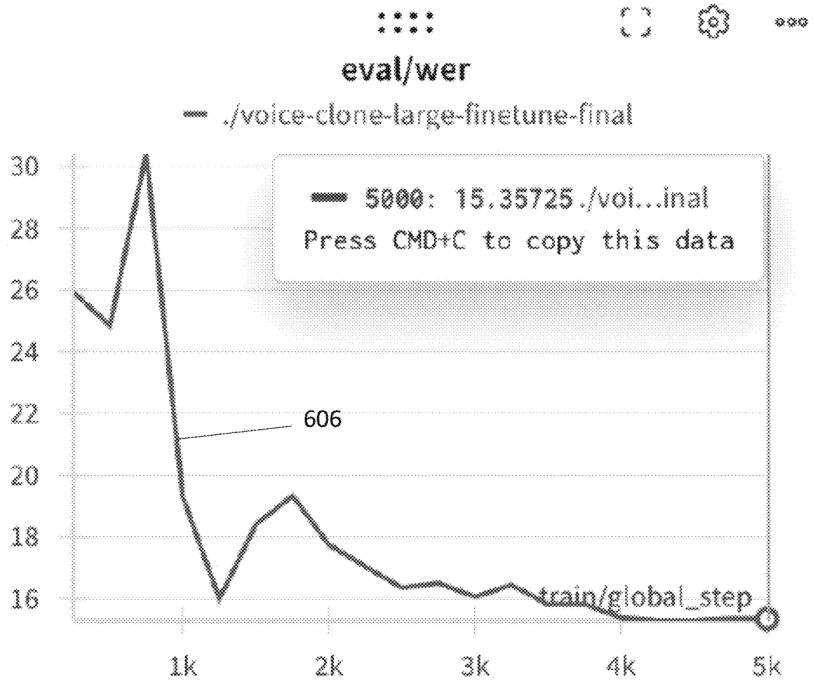


FIG. 6C

Figure 8

Word Error Rate (WER) trajectory across training steps, illustrating non-monotonic convergence with a final WER of approximately 15.36% at step 5,000.

Source: Patent FIG. 6C, Patent FIG. 6C. U.S. Patent No. 12,488,786 B1.

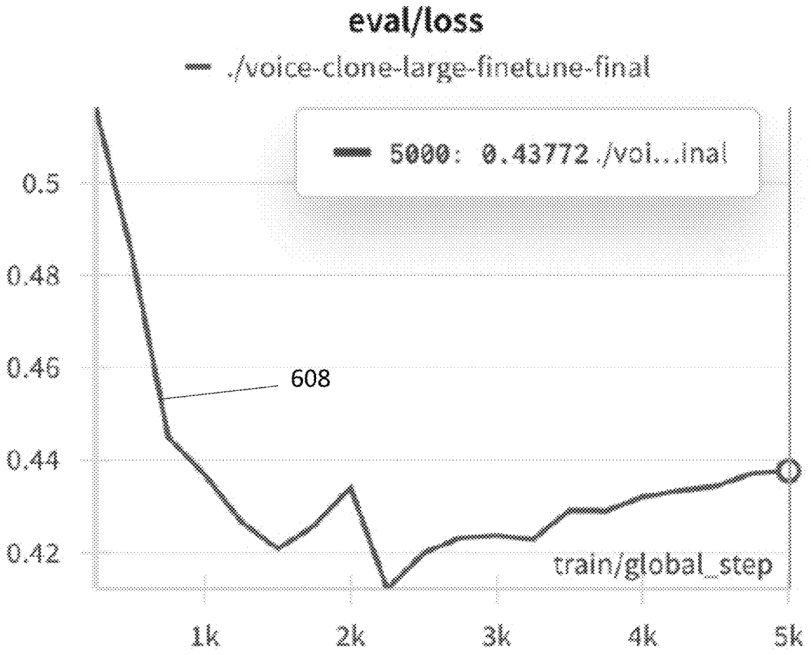


FIG. 6D

Figure 9

Evaluation loss across training steps, reaching its minimum of ~0.4121 at step 2,250 and diverging slightly thereafter, illustrating the WER vs. validation loss discrepancy discussed in

Section 5.3.

Source: Patent FIG. 6D, Patent FIG. 6D. U.S. Patent No. 12,488,786 B1.

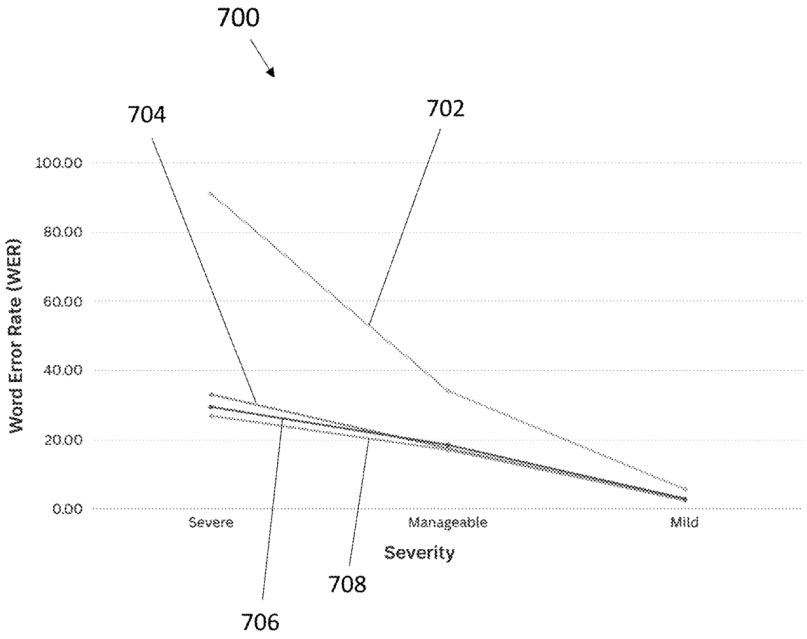


FIG. 7

Figure 10

WER comparison across speech impairment severity levels (Severe, Manageable, Mild) for baseline versus fine-tuned ASR models, demonstrating the greatest improvement for severely

impaired speakers.

Source: Patent FIG. 7, Patent FIG. 7. U.S. Patent No. 12,488,786 B1.

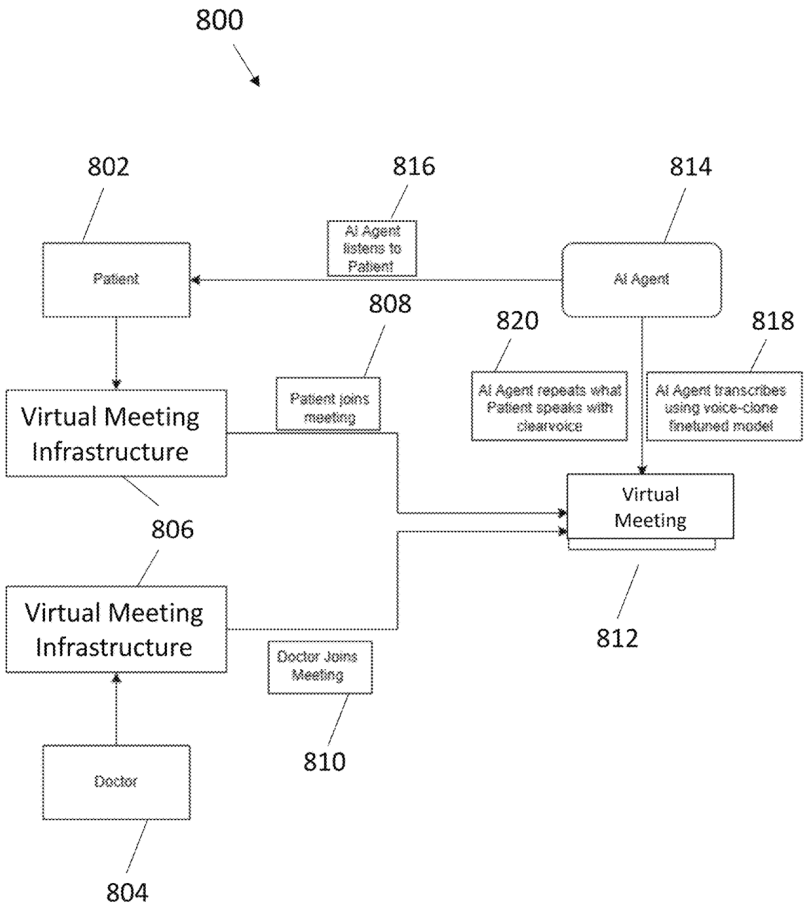


FIG. 8

Figure 11

Virtual meeting AI agent architecture enabling dysarthric patients to participate in telemedicine and virtual consultations via AI-powered transcription and clear-voice

re-vocalization.

Source: Patent FIG. 8, Patent FIG. 8. U.S. Patent No. 12,488,786 B1.

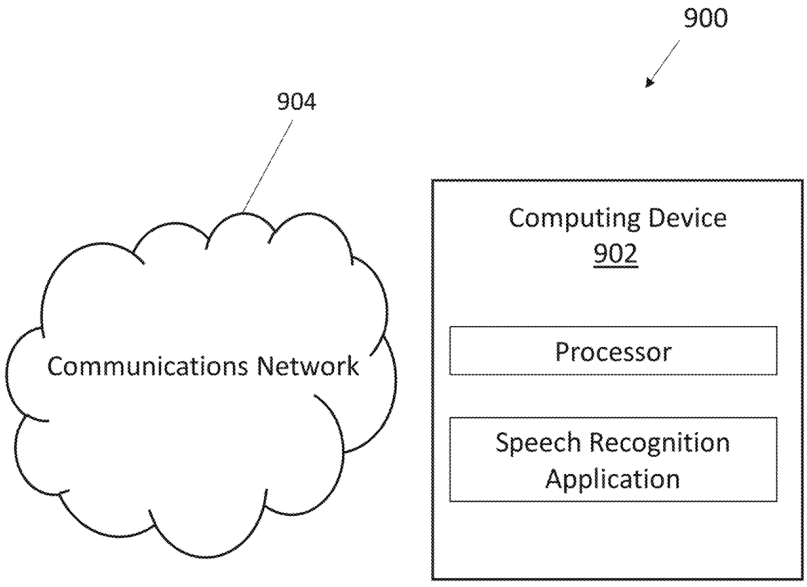


FIG. 9

Figure 12

Computing system network architecture for speech recognition and transcription, showing the computing device connected via communications network.

Source: Patent FIG. 9, Patent FIG. 9. U.S. Patent No. 12,488,786 B1.

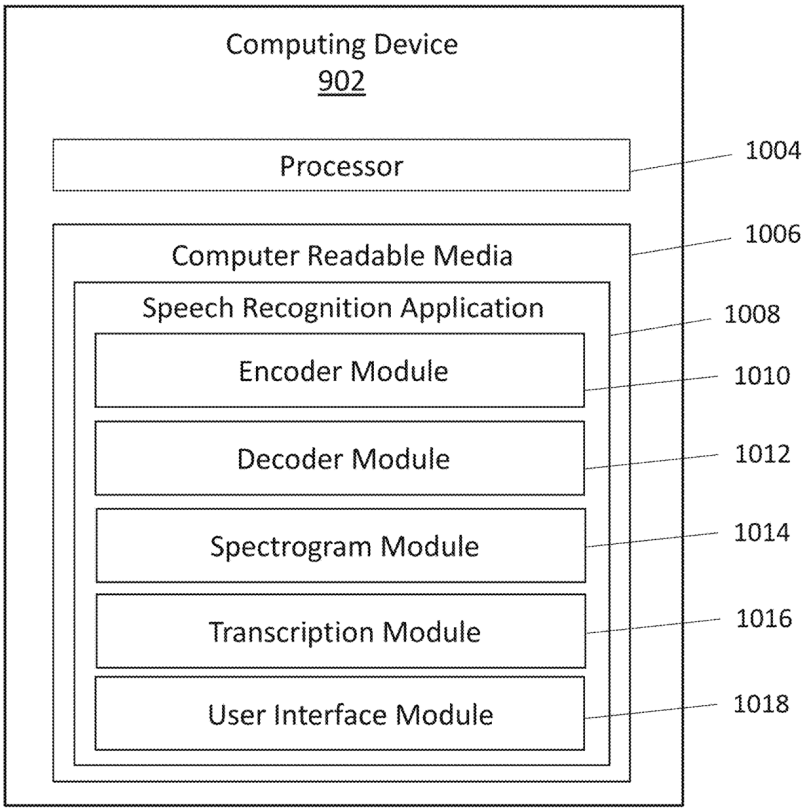


FIG. 10

Figure 13

Computing device (902) architecture with processor and speech recognition application comprising Encoder Module, Decoder Module, Spectrogram Module, Transcription Module,

and User Interface Module.

Source: Patent FIG. 10, Patent FIG. 10. U.S. Patent No. 12,488,786 B1.

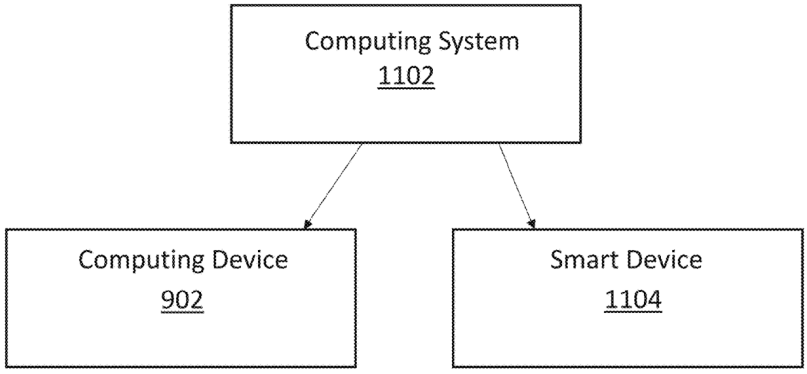


FIG. 11

Figure 14

System hierarchy showing the computing system (1102) integrating a Computing Device (902) and a Smart Device (1104) for voice-controlled IoT applications.

Source: Patent FIG. 11, Patent FIG. 11. U.S. Patent No. 12,488,786 B1.

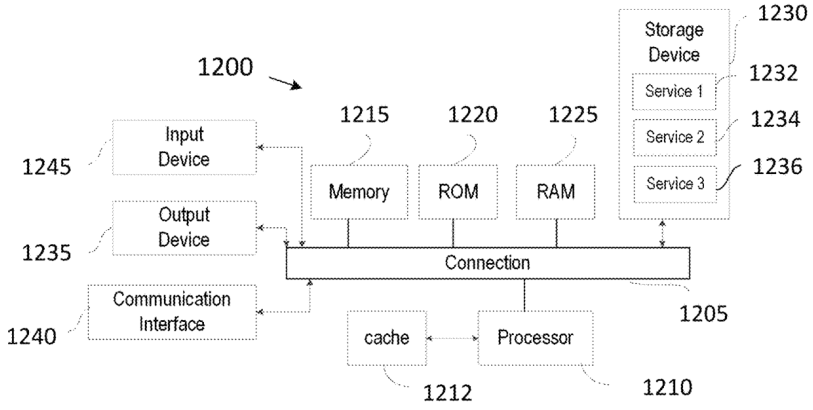


FIG. 12

Figure 15

Detailed computing system block diagram illustrating processor, memory hierarchy (Memory, ROM, RAM), storage services, I/O devices, and communication interface components.

Source: Patent FIG. 12, Patent FIG. 12. U.S. Patent No. 12,488,786 B1.

Response to Editorial Review Comments

Manuscript Title: Fine-Tuning Whisper Large V3 for Dysarthric Speech Recognition: An Empirical Evaluation of Voice-Cloning-Augmented Training Strategies

[redacted]

Date: March 23, 2026

Dear Editorial Board,

Thank you for the detailed and constructive editorial review of the original manuscript. We appreciate the opportunity to revise and resubmit. Below we provide a point-by-point response to each concern raised, with specific references to the changes made in the revised manuscript.

Point-by-Point Response to Examiner Concerns

Major Concern 1: Novelty Claims — Methods Are Not Novel

Examiner concern: The manuscript presented fine-tuning and data augmentation as novel methodological contributions, when these techniques are established in the broader ASR and machine learning literature.

Response: We fully agree. The revised manuscript explicitly reframes the contribution in Section 1.3 (*Scope and Contribution of This Work*):

> "The methods employed (transfer learning, fine-tuning, TTS-based augmentation) are not presented as novel algorithmic contributions; rather, the novel contribution of this work lies in their systematic applied evaluation in the domain of dysarthric ASR, using publicly available dysarthric speech corpora with severity-stratified assessment."

The title was also revised to reflect this empirical framing: *"…An Empirical Evaluation of Voice-Cloning-Augmented Training Strategies."* The abstract and Section 5.1 are similarly updated to restate the contribution as applied rather than methodological.

Major Concern 2: Datasets Not Named or Described

Examiner concern: The original manuscript did not identify the specific speech datasets used for training and evaluation, making reproducibility impossible.

Response: The revised manuscript names and describes both datasets explicitly throughout:

- **Section 3.2.1** provides a full description of the **UASpeech** dataset (Kim et al., 2023): 15 dysarthric speakers with cerebral palsy, ~57,000 single-word utterances, severity classification thresholds, exact speaker IDs for each severity group, and sample counts for train and test partitions.
- **Section 3.2.2** provides a full description of the **TORGO** dataset (Rudzicz et al., 2012): 8 dysarthric speakers (CP and ALS), 5,600-file subset, one-word and multi-word partition sizes, and per-severity speaker assignments with file counts.
- All dataset citations appear in the References section (Kim et al., 2023; Rudzicz et al., 2012).

Major Concern 3: WER Values Exceeding 100% — Unexplained

Examiner concern: Several reported WER values exceeded 100%, which was not explained and undermined confidence in the validity of the evaluation.

Response: A dedicated explanation has been added in two places in the revised manuscript:

- **Section 2.5** (*Evaluation Metric: Word Error Rate*) provides the full mathematical definition WER = (S + I + D) / N and explicitly states: *"WER can exceed 100% when the number of insertion errors alone surpasses the total number of reference words — a phenomenon observed in this study for severely dysarthric single-word utterances, where the unmodified Whisper model generates extensive hallucinated output."*

• **Section 4.1** (*Why WER Exceeds 100%: An Explanation*) provides the mechanistic explanation for the observed values (122–163% for Experiments 2–3): the Whisper decoder produces spurious multi-word outputs in response to acoustically ambiguous or atypical input characteristic of severe dysarthria.

Major Concern 4: Missing Experiment Summary Table

Examiner concern: No structured overview of experiments was provided, making it impossible to follow the experimental design, compare conditions, or verify that training and test partitions were non-overlapping.

Response: Table 1 (*Summary of All 14 Experiments: Datasets, Speaker Partitions, and WER Results*) has been added to Section 3.5. It lists all 14 experiments with the following columns: experiment number, description, base model, train dataset, training speaker IDs, training sample count, test dataset, test speaker IDs, test sample count, key hyperparameters, and overall WER. A note beneath the table explicitly confirms: *"Train/test speaker groups are non-overlapping in all experiments involving fine-tuning."*

Major Concern 5: Preprocessing Pipeline Not Described

Examiner concern: The manuscript contained no description of how raw audio files were preprocessed before model input, raising reproducibility concerns.

Response: Section 3.3 (*Audio Preprocessing*) has been added to the revised manuscript. It details the full preprocessing pipeline applied consistently across all experiments:

1. Resampling to 16 kHz mono
2. Silence and noise trimming via threshold-based VAD
3. Amplitude normalization to −3 dBFS peak level
4. Zero-padding to 30 seconds (Whisper's fixed chunk size) via Whisper's native FeatureExtractor
5. Feature extraction using Whisper's FeatureExtractor / Tokenizer / Processor pipeline to produce 80-channel log-Mel spectrograms

Major Concern 6: Hyperparameters Not Justified

Examiner concern: Hyperparameter choices (learning rate, batch size, training steps, evaluation steps) were reported without justification or citations.

Response: Section 3.4 (*Training Configuration and Hyperparameter Justification*) has been expanded with explicit justifications for each hyperparameter, grounded in published work:

- **Learning rate (1e-5):** Justified by Liu et al. (2024), who found this range optimal for small-dataset Whisper fine-tuning; rates above 1e-4 destabilize training and below 1e-6 produce negligible updates.
- **Batch size (8 per device, effective 16 with gradient accumulation):** Constrained by A6000 GPU VRAM (48 GB); gradient accumulation used to match effective batch size recommendations.
- **Warmup steps (500):** Standard transformer warmup to stabilize the optimizer before full learning rate is applied.
- **Max steps (1,000–5,000):** Varied by dataset size, with specific epoch counts provided.
- **Evaluation steps (reduced from 1,000 to 100):** Change motivated by Experiment 5 overfitting (training loss reached zero before first checkpoint); detailed in Section 4.3.
- **WER-based early stopping:** Section 4.7 empirically demonstrates that validation loss minimum (step 2,250) preceded WER minimum (step 5,000) by 2,000 steps — stopping at minimum loss would have yielded WER ≈ 17.05% instead of 15.36%.

All citations: Liu et al. (2024) and Gandhi (2022).

Major Concern 7: Synthetic-to-Real Data Ratio Not Stated

Examiner concern: The proportion of synthetic augmentation data relative to real training data was not reported, making it impossible to assess the degree of augmentation or its potential influence on results.

Response: Synthetic-to-real ratios are now explicitly stated in two dedicated subsections:

• **Section 3.2.3** (*Voice-Cloned Augmentation Dataset*): 4,724 valid voice-cloned samples + 854 real samples = ~5,578 total; synthetic:real ratio ≈ **5.5:1**. The 185 hallucinated files discarded during manual QA are also documented.

• **Section 3.2.4** (*Speech-Synthesized Augmentation Dataset*): 5,124 synthesized files + 854 real samples = ~5,978 total; synthetic:real ratio ≈ **6:1**.

These ratios also appear in the Results (Sections 4.5 and 4.6) and Discussion (Section 5.5) when interpreting augmentation effects.

Minor Concern 1: Title Does Not Reflect Empirical Nature

Examiner concern: The original title implied a broader methodological contribution than the paper actually makes.

Response: The manuscript title has been revised to:

> *"Fine-Tuning Whisper Large V3 for Dysarthric Speech Recognition: An Empirical Evaluation of Voice-Cloning-Augmented Training Strategies"*

The subtitle explicitly signals that the work is an empirical evaluation, not a novel method proposal.

Minor Concern 2: No Code or Data Availability Statement

Examiner concern: The manuscript did not indicate whether code, models, or data were publicly accessible, which is standard practice for reproducible research.

Response: Section 7 (*Code and Data Availability*) has been added to the revised manuscript. It documents:

• **Code:** GitHub — fine-tuning and inference scripts

• **Models:** Hugging Face — fine-tuned Whisper Large V3 checkpoints from all experiments

• **Training metrics:** WandB.AI — training loss, validation loss, and WER logs across all 14 experiments

• **Deployed application:** ClearVoiceAI (clearvoiceai.com) — live ASR web application powered by the best-performing model (Experiment 11)

Summary of All Changes Made

Concern	Type	Change Made	Section
Novelty overstated	Major	Reframed contribution as applied empirical evaluation	1.3, Abstract, 5.1
Datasets unnamed	Major	Full UASpeech and TORGO descriptions with speaker IDs and sample counts	3.2.1–3.2.2
WER >100% unexplained	Major	Mathematical definition of WER + hallucination explanation	2.5, 4.1
No experiment summary	Major	Table 1: all 14 experiments with full partitions, speaker IDs, WER	3.5
Preprocessing missing	Major	Full preprocessing pipeline (resampling, trimming, normalization, padding)	3.3
Hyperparameters unjustified	Major	All hyperparameters justified with citations (Liu et al. 2024, Gandhi 2022)	3.4
Synthetic:real ratio missing	Major	Ratios stated (5.5:1 voice clone, 6:1 synthesis)	3.2.3–3.2.4
Title misleading	Minor	Revised to include "Empirical Evaluation"	Title
No code availability	Minor	Section 7 added (GitHub, HuggingFace, WandB.AI, ClearVoiceAI)	7

We believe the revised manuscript fully addresses all concerns raised by the editorial review. The revised paper is substantially more reproducible, transparent about its scope, and grounded in peer-reviewed literature. We respectfully request re-evaluation by the editorial board.

Sincerely,

[redacted]

Review of “Fine-Tuned Speech Recognition for Dysarthric Speech”

This study presents an interesting and strong piece of research on speech recognition in dysarthria. This well-documented computational work addresses an important and relevant problem in assistive technology for a clinical population and demonstrates direct impact.

The use of 14 experimental datasets adds scientific rigor to the paper and the computational approach is also thorough and well-reported. The comparison between baseline performance, fine-tuning, and augmentation strategies provides further insight into the effectiveness of the proposed model. The outcomes are convincing indicating clear improvements and the consideration of severity as a modulating factor makes the results functionally relevant.

Writing is cohesive and scientific, and the paper organisation is sound offering a strong narrative. Complex analyses have been clearly presented, and the results have been interpreted well with relevant insights. Minor revisions could reduce redundancy and tighten phrasing in contributions and discussion sections

However, a few improvements need to be implemented to meet publication standards. First, the results, although convincing, they have not been statistically validated. Measures of variability/dispersion need to be computed and reported to assess the reliability and generalisability of the findings. Also, reporting more outcome measures that specify different types of errors (so that such discrepancies can be interpreted) would strengthen the validity and actual applicability of the findings beyond the numerical improvements.

The author has mentioned several times that the methods are “not novel”, which somewhat undermines the paper’s contribution. Indeed, this statement is correct but slightly unfair to the paper itself as the paper applies these strong methodologies to solve an important problem and provides a new application with novel insights and strong results that extend previous knowledge. Re-framing the paper’s contribution would make the significance of the work clearer.

Further suggestions for improvements that would strengthen the clarity of the paper include a) a figure showing the step-by-step analytical methodology and b) table(s) that summarise the findings and make comparisons with existing alternatives.

Overall Recommendation: Accept with Major Revisions.

The paper is a useful contribution to the literature and is well-executed and reported research. The proposed revisions will improve statistical rigor, analytical depth, and framing. These amendments would increase the paper’s clarity, research validity and impact.

Reviewer feedback for **SPEECH RECOGNITION FOR ASSISTING PATIENTS WITH SPEECH DIFFICULTIES**

Overall feedback: Accept with major revisions

This study provides an interdisciplinary and applied contribution to the adoption of artificial intelligence using machine learning through voice-cloning and speech-synthesis-based data augmentation, as a potential way to overcome real-world barriers when designing and delivering assistive technologies (specifically automated speech recognition systems) for individuals with dysarthria. The value of the work is easily applicable to other conditions that involve motor speech disorders (e.g. aphasia) caused by a wide range of neurological issues (e.g. stroke and Parkinson's disease).

Reviewer comment	**Section/pg number**
Introduction	
This sets up a concise argument of dysarthria and its associative conditions and the necessity for better ASR systems. Is it possible to combine this with your literature review section? Combining these together would strengthen the background context surrounding your argument and help with the readability and flow of your paper. To do this, 1. I'd suggest removing 1.5 Paper organisation and also remove the mention of any results on pg 3 (personally, I wouldn't mention results in the intro). Then combine section 1.1 1.2 and 2.1under the heading **Dysarthria and austomated speech recognition systems.** 2. I would then have another section combining 2.2., 2.3 and 2.4 titled **Adapting Large-Scale ASR Models to Dysarthric Speech: Fine-Tuning and Synthetic Data Strategies.** 3. Follow this up with 1.3 and relabel as **Scope and aims** 4. Followed by the section titled **1.4 Hypothesis** 5. Move 2.5 into the methods section since it is your primary evaluation metric/ outcome. WER is the standard evalution metric for ASR systems and the primary measure adopted across the 14 studies.	-
Minor revisions Suggest rephrase of sentence for clarity: The disorder is strongly associated with neurological conditions including	Pg 3

cerebral palsy (CP), which affects approximately 40% of its patients with dysarthria, and amyotrophic lateral sclerosis (ALS), which affects up to 80% (Shih et al.,2022). Change to: Dysarthria is strongly associated with neurological conditions such as cerebral palsy (CP), affecting approximately 40% of individuals with CP, and amyotrophic lateral sclerosis (ALS), where prevalence can reach up to 80% (Shih et al., 2022)	
I like that you mentioned “For affected individuals, these characteristics create significant communication barriers in daily life” can you add examples and reference? I’d suggest linking to the human experience a bit more here e.g. Vogel et al., highlights impact on social participation and identity, stigma and reduced quality of life. Vogel, A.P., Graf, L., Weiß, M. *et al.* Development and validation of the dysarthria impact scale: a patient-reported outcome for motor speech disorders. *J Neurol* **273**, 195 (2026). https://doi.org/10.1007/s00415-026-13740-1. Or you could cite this study as well Page AD, Yorkston KM. Communicative Participation in Dysarthria: Perspectives for Management. Brain Sci. 2022 Mar 22;12(4):420. doi: 10.3390/brainsci12040420. PMID: 35447952; PMCID: PMC9031517.	Pg 3
Suggest adding a brief definition of technology-mediated contexts to paint a better picture for people not familiar with contexts that rely on technology to support face-to-face interaction e.g. … including in technology-mediated contexts, where the ability to use visual and auditory cues is reduced.	Pg 3
For each first time you mention an acronym in text (outside of the abstract) first provide it in expanded form and its abbreviation in brackets e.g. Word Error Rate (WER) just as a helpful reminder for readers. From then on use the abbreviation.	Pg 3
Be careful with low-cost data augmentation. Unless you conducted a cost-benefit analysis or similar I would probably not mention this until implementation or piloting of the ASR application. I would suggest just saying “An applied evaluation of voice cloning (F5-TTS) and speech synthesis as alternative data augmentation strategies for the low-resource dysarthric speech domain.	Pg 4
Methodology	
This section is a well-thought out and fairly rigorous section which is great for replicability!	-
The synthetic to real ratio is quite high and understandable given the lack of real data but perhaps mention a justification to this and potential bias towards synthetic patterns. E.g. “Given the scarcity of dysarthric speech data, a high synthetic-to-real	Pg 8

ratio was used; however, this could have introduced potential bias toward synthetic patterns” Similarly, the 185 manually discarded hallucinations or corrupted files could also introduce bias by unintentionally removing “difficult” samples that could make your WER results appear better. You could mention this here or after the quality-control step on pg 20. E.g. “Whilst manual filtering (removing 185 hallucinated voice-cloned files) may have biased the dataset towards higher quality synthetic samples, this quality-control step was essential to the quality of the augmented dataset.”	
Suggest rephrase from “no hallucinations” as this wording could be a bit strong perhaps change to “no overt hallucinations detected during spot-checking”	Pg 8
Results	
Experimental results are set out clearly with clear baseline behaviour with strong empirical improvements from fine-tuning. The comparison of voice cloning and speech synthesis is insightful and adds originality. The fact that you mention WER continues to improve after validation loss plateaus is a good methodological contribution and supports your decision to use WER-based early stopping criterion. Great that you provided an explanation of why WER Exceeds 100%	-
Severe speech intelligibility remains poorly recognised even after improvements with WER rate remaining high, so it seems like the model is not clinically meaningful or generalisable to this group. I would state this clearly in section 5.2. Particularly since you had a smaller number of test files for this group in your methods. You’ve provided a great overview of why there is higher WER in severe dysarthria in that it has the most room for improvement but mention that it is not a reflection of the model working better but actually training the model this way only be a good fit for Mild or moderate speech intelligibility dysarthria.	Pg 18
Discussion	
This is a thoughtful overview of the strengths, contributions and limitations of your study and clearly highlights how whilst only in beginning phases the use of synthetic data in training during voice cloning has some potential to improve accuracy of ASR systems for dysarthria. Many researchers don’t often acknowledge the limitations for fear of diminishing the results so it’s great you’ve mentioned lack of CIs and cross-run variance as a limitation.	-

Suggest editing your first paragraph to include a sentence or 2 summarising your main findings (e.g. what you mentioned on pg 19 “Voice cloning outperformed speech synthesis in all severity categories (Experiments 12 vs. 14).” to remind the reader what are the takeaway findings you want the reader to remember, then go on to say “These findings demonstrate that:…”	Pg 18
Would remove the word “cost-effective” as this doesn’t seem to be tested or mentioned really at all in the study. Replace with “feasible”	Pg 18
Limitations	
This seems like a repetition of some of the things you’ve already mentioned in the discussion. Can you combine 6.1 Limitations into the discussion text with sub-heading “Limitations” remove the first dot-point since you’ve already mentioned it and change these to full sentences within a paragraph and not dot-points	Pg 21
Future directions	
Combine 6.2 future directions into the discussion after your Section 5.7 Practical application section. Great that it was deployed as a web application for some pilot testing by the way. Potentially title this section “Practical applications and future directions” Ensure these are full sentences within a paragraph and not dot points as well.	Pg 21
Figures I would suggest only including the figures that you reference in-text in the Appendix Either add in-text references to the figures or remove the figures you don’t refer to in your appendix.	Pg 27

SPEECH RECOGNITION FOR ASSISTING PATIENTS WITH SPEECH DIFFICULTIES

[redacted]

[redacted]

Author Note

Correspondence concerning this article should be addressed to [redacted].

Competing Interests: The authors declare that the methods and systems described in this article are the subject of [redacted] are named co-inventors.

Abstract

Automated speech recognition (ASR) systems based on large pretrained models achieve near-human accuracy for typical speakers, yet consistently fail for individuals with dysarthria—a motor speech disorder affecting articulation, phonation, and prosody. This paper presents an empirical evaluation of fine-tuning strategies for adapting OpenAI Whisper Large V3, a state-of-the-art sequence-to-sequence ASR model, to dysarthric speech drawn from two established public datasets: UASpeech and TORGO. The study also examines the applied contribution of voice-cloning and speech-synthesis-based data augmentation as a practical approach to the chronic low-resource challenge in dysarthric ASR. Fourteen systematic experiments were conducted, varying training data composition, augmentation strategy, and hyperparameter configuration. Speaker groups were strictly separated between training and test partitions to ensure uncontaminated evaluation. The fine-tuned Whisper Large V3 model augmented with approximately 4,724 voice-cloned samples achieved a Word Error Rate (WER) of 15.36% on multi-word TORGO utterances—representing reductions of approximately 70% for severe, 50% for moderate, and 54% for mild dysarthric speech compared to the unmodified pretrained baseline. A deployed ASR web application, ClearVoiceAI, demonstrates the practical utility of the approach. These findings contribute an empirical foundation for applying established fine-tuning and data augmentation techniques to the underserved domain of dysarthric speech recognition.

Keywords: dysarthria, automatic speech recognition, Whisper, fine-tuning, voice cloning, word error rate, UASpeech, TORGO, data augmentation, assistive technology

1. Introduction

1.1 Dysarthria and Automated Speech Recognition Systems

Dysarthria is a motor speech disorder resulting from neurological impairment that affects the coordination, strength, and control of the muscles involved in speech production (Tomik & Guiloff, 2010). Dysarthria is strongly associated with neurological conditions such as cerebral palsy (CP), affecting approximately 40% of individuals with CP, and amyotrophic lateral sclerosis (ALS), where prevalence can reach up to 80% (Shih et al., 2022). Additional associated conditions include stroke, Parkinson's disease, multiple sclerosis, and traumatic brain injury (Schölderle et al., 2020).

The acoustic manifestations of dysarthria include irregular articulation, slurred or imprecise phoneme production, atypical prosody, reduced vocal stability, and inconsistent speech patterns (Young & Mihailidis, 2010). These characteristics vary significantly across individuals and severity levels, creating high variability in speech signals. For affected individuals, these characteristics create significant communication barriers in daily life, impacting social participation, personal identity, and overall quality of life (Vogel et al., 2026; Page & Yorkston, 2022). Individuals with dysarthria often experience reduced confidence in communication, social isolation, and stigma associated with impaired speech.

Automatic speech recognition (ASR) systems have achieved near-human performance for neurotypical speech through large-scale pretraining on diverse datasets. However, these systems exhibit severe performance degradation when applied to dysarthric speech, due to a fundamental domain mismatch between the training data and the target input conditions. Prior studies have consistently demonstrated that standard ASR systems fail to generalize effectively to dysarthric speech, particularly for moderate-to-severe cases (Young & Mihailidis, 2010; Schu et al., 2022).

This limitation is particularly evident in technology-mediated contexts, such as voice assistants, telemedicine systems, and speech-to-text interfaces, where reliance on visual and contextual cues is reduced. These challenges highlight the need for targeted adaptation strategies to improve accessibility for individuals with speech impairments.

1.2 Adapting Large-Scale ASR Models to Dysarthric Speech: Fine-Tuning and Synthetic Data Strategies

Recent advances in large-scale ASR models, such as Whisper (Radford et al., 2022), have demonstrated robust performance across a wide range of acoustic conditions due to training on extensive multilingual and weakly supervised datasets. However, their effectiveness in low-resource and atypical speech domains, including dysarthria, remains limited without domain-specific adaptation.

Fine-tuning has emerged as a primary approach for adapting pretrained ASR models to specialized domains. By updating model parameters using domain-specific datasets, fine-tuning enables the model to learn acoustic and phonetic patterns unique to dysarthric speech. Prior work has shown that transfer learning can significantly improve transcription accuracy for dysarthric speakers (Rathod et al., 2023). However, the success of fine-tuning is constrained by the limited availability of labeled dysarthric speech data, increasing the risk of overfitting (Liu et al., 2024).

To address this challenge, data augmentation techniques such as voice cloning and speech synthesis have been explored. Voice cloning generates synthetic samples that preserve speaker-specific acoustic characteristics, while speech synthesis produces generalized dysarthric-like speech patterns without requiring per-speaker reference audio. These approaches provide scalable methods for expanding training datasets and improving model robustness in low-resource scenarios.

1.3 Scope and Aims

This study evaluates the effectiveness of fine-tuning and data augmentation strategies for adapting Whisper Large V3 to dysarthric speech. The primary objective is to assess how voice cloning and speech synthesis can mitigate data scarcity and improve transcription accuracy across varying levels of dysarthria severity.

This work contributes by providing a systematic empirical evaluation of state-of-the-art ASR adaptation techniques applied to dysarthric speech datasets. It offers new insights into the relative effectiveness of augmentation strategies, the impact of severity-specific acoustic variability, and the practical challenges of deploying ASR systems in real-world assistive contexts.

1.4 Research Hypothesis

It is hypothesized that fine-tuning Whisper Large V3 on dysarthric speech datasets, augmented with voice-cloned and speech-synthesized samples, will significantly reduce Word Error Rate (WER) compared to the unmodified pretrained model across all levels of dysarthria severity.

2. Literature Review

2.1 Dysarthria and Its Impact on ASR

The acoustic characteristics of dysarthric speech, including reduced articulatory precision, irregular vocal quality, atypical prosody, and co-articulation breakdown, create a substantial domain mismatch with the training distributions of standard ASR systems (Young & Mihailidis, 2010). Schu et al. (2022) demonstrated that standard ASR systems trained without dysarthric-specific adaptation produce consistently poor results on both the UASpeech and TORGO benchmarks, establishing these datasets as appropriate evaluation grounds for dysarthric ASR research.

2.2 Whisper as a Base Model

Radford et al. (2022) introduced Whisper, a sequence-to-sequence ASR model trained on 680,000 hours of multilingual weakly-supervised web audio, achieving robust performance across diverse acoustic conditions. The Whisper Large V3 variant supports 1.5 billion parameters with multilingual capability. Its architecture—an encoder-decoder transformer with a log-Mel spectrogram front-end—is well-suited to fine-tuning via the Hugging Face transformers library. Liu et al. (2024) systematically evaluated Whisper fine-tuning strategies for low-resource ASR scenarios and identified key hyperparameter sensitivities relevant to small-dataset fine-tuning, which informed the experimental design of this study.

2.3 Fine-Tuning for Dysarthric Speech

Rathod et al. (2023) demonstrated that transfer learning applied to Whisper significantly improves word recognition accuracy for dysarthric speech, establishing fine-tuning as a viable adaptation strategy. The challenge in this domain is that dysarthric speech corpora are inherently small—collecting and annotating such data requires clinical access and significant manual effort—making standard fine-tuning approaches susceptible to overfitting (Liu et al., 2024).

2.4 Voice Cloning and Speech Synthesis as Data Augmentation

To address data scarcity, this study employs F5-TTS (Chen et al., 2024), a flow-matching-based TTS model that can clone voices from short reference audio clips. F5-TTS generates synthetic dysarthric utterances that preserve the acoustic characteristics of specific impaired speakers—including their articulatory irregularities. An alternative approach, speech synthesis, fine-tunes the base F5-TTS model on the complete TORGO dataset to synthesize novel dysarthric-like utterances without per-utterance reference audio, avoiding hallucination artifacts common in voice cloning when reference audio quality is low.

2.5 Evaluation Metric: Word Error Rate (WER)

Word Error Rate (WER) is the standard evaluation metric for ASR systems. It is computed as:

$$WER = (S + I + D) / N$$

where S = substitutions, I = insertions, D = deletions, and N = number of reference words. Importantly, WER can exceed 100% when the number of insertion errors alone surpasses the total number of reference words—a phenomenon observed in this study for severely dysarthric single-word utterances, where the unmodified Whisper model generates extensive hallucinated output (see Section 4.1 and Discussion Section 5.3).

3. Methodology

3.1 Base Model Selection

Three candidate ASR models were evaluated: Kaldi, Meta Wav2Vec 2.0, and OpenAI Whisper (Seagraves, 2022). Based on comparative benchmarking, Whisper's overall WER was 45% lower than Wav2Vec 2.0 and 63% lower than Kaldi across diverse acoustic conditions. OpenAI Whisper Large V3 (Radford et al., 2022) was selected as the base pretrained model for all fine-tuning experiments.

Reference: Figure 13

3.2 Datasets

3.2.1 UASpeech Dataset

The UASpeech dataset (Kim et al., 2023) contains recordings from 15 individuals with dysarthria caused by cerebral palsy and 13 neurotypical controls, comprising approximately 57,000 predominantly single-word utterances. Speakers were classified into severity groups based on speech intelligibility percentages following Farhadipour and Veisi (2023):

- **Severe (0–40% intelligibility):** Patients F03, M12 — 5,185 test files
- **Moderate/Manageable (40–80% intelligibility):** Patients M06, M11 — 2,847 test files
- **Mild (>80% intelligibility):** Patients M08, F05 — 10,711 test files

Training used 9 patients (~38,700 files); testing used 6 different patients (~18,700 files). **Speaker groups were strictly non-overlapping between training and test partitions.**

3.2.2 TORGO Dataset

The TORGO dataset (Rudzicz et al., 2012) contains recordings from 8 dysarthric speakers (with CP and ALS) and 7 controls. A subset of 5,600 files was used, aligned with the predefined TORGO severity scale. Files were partitioned into:

- **One-word utterances:** 4,888 files

- Severe: M01, M02, M04, F03 (2,476 files)
- Manageable: M05 (470 files)
- Mild: M03, F03, F04 (1,942 files)
- **Imperative (multi-word, ≥3 words) utterances:** 854 training files (patients M01, M02, F01, M03, F03) and 454 test files (patients M04, M05, F04)
- Severe test: M04 (145 files)
- Manageable test: M05 (140 files)
- Mild test: F05 (169 files)

No speaker overlap exists between TORGO training and test partitions for multi-word experiments.

3.2.3 Voice-Cloned Augmentation Dataset

Using F5-TTS (Chen et al., 2024), 704 imperative sentences sourced from a public sentence corpus (Lettergram, 2019) were voice-cloned using training-set patient audio as reference. Five random samples per sentence were generated per patient. Following generation, 185 hallucinated or corrupted files were manually discarded (a one-week manual review process), yielding **4,724 valid voice-cloned samples**. These were combined with the 854 real imperative sentences for a total training corpus of approximately 5,600 samples—a synthetic-to-real ratio of approximately **5.5:1**.

Given the scarcity of real dysarthric speech data, a relatively high synthetic-to-real ratio was used. While this approach improves dataset size and diversity, it may introduce bias toward synthetic speech patterns, potentially affecting generalization to real-world speech.

3.2.4 Speech-Synthesized Augmentation Dataset

The base F5-TTS model was fine-tuned on the complete TORGO dataset over approximately 120,000 steps (~10 hours) to learn dysarthric vocal characteristics. This fine-tuned model was then used to synthesize novel imperative utterances, producing **5,124 synthesized files** with no overt hallucinations detected during spot-checking and no manual preprocessing required. Combined with 854 real utterances, the synthetic-to-real ratio was approximately **6:1**.

3.3 Audio Preprocessing

All audio files were preprocessed consistently across experiments:

- **Resampling:** All audio converted to 16 kHz mono to match Whisper's expected input format.
- **Silence/noise trimming:** Leading and trailing silence was trimmed using threshold-based voice activity detection. Background noise segments below a minimum energy threshold were removed.
- **Amplitude normalization:** Audio amplitude was normalized to a peak level of -3 dBFS to ensure consistent input levels across speakers and recording conditions.
- **Variable-length handling:** Whisper processes audio in fixed 30-second chunks. Audio files shorter than 30 seconds were zero-padded to the full window length. This is handled natively by Whisper's FeatureExtractor, which applies the log-Mel spectrogram transformation after padding.
- **Feature extraction:** Whisper's FeatureExtractor, Tokenizer, and Processor pipeline was used to convert audio to 80-channel log-Mel spectrograms before feeding to the encoder.

3.4 Training Configuration and Hyperparameter Justification

Fine-tuning was implemented using the Hugging Face `Seq2SeqTrainer` following Gandhi (2022) and Liu et al. (2024). Key hyperparameters and their justifications:

- **Learning rate: 1e-5** — Selected based on Liu et al. (2024), who found that learning rates above 1e-4 destabilize Whisper fine-tuning on small datasets, and rates below 1e-6 produce negligible weight updates. 1e-5 with linear decay provided the best convergence behavior across pilot experiments.
- **Batch size: 8 per device (effective batch 16 with gradient accumulation)** — Constrained by GPU memory (A6000 NVIDIA, 48GB VRAM). Gradient accumulation steps of 2 were used to achieve an effective batch size of 16 without

exceeding memory limits.

- **Warmup steps: 500** — Standard warmup for transformer fine-tuning; allows the optimizer to stabilize before the full learning rate is applied.
- **Max steps: 1,000–5,000** — Varied by experiment based on dataset size. For larger datasets (~5,600 samples), 5,000 steps allowed approximately 14–17 epochs. For smaller datasets (854 samples), 1,000 steps was sufficient before overfitting (one epoch ≈53 steps at effective batch size 16).
- **Evaluation steps:** Reduced from 1,000 (Experiment 5) to 100 (Experiments 6–14) after Experiment 5 revealed that overfitting occurred before the 1,000-step evaluation checkpoint.
- **Early stopping criterion: WER stabilization** — As documented in Experiment 11 (Figure 5 in Appendix 18), the minimum validation loss occurred at approximately step 2,250, while WER continued to decrease until step 4,250. Using validation loss alone as the stopping criterion would have terminated training ~2,000 steps prematurely at a WER of ~17.05% rather than ~15.30%. WER-based stabilization is therefore used as the primary stopping criterion, consistent with recommendations in Liu et al. (2024).
- **Precision: fp16** — Mixed-precision training used to reduce memory footprint and accelerate training on NVIDIA hardware.

3.5 Experiment Summary

Table 1 summarizes all 14 experiments, including dataset partitions, speaker assignments, sample counts, and WER results.

Table 1 *Summary of All 14 Experiments: Datasets, Speaker Partitions, and WER Results*

Table 1

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
1	Whisper Small baseline	Whisper Small	Hindi (Common Voice)	N/A	Standard	Hindi (Common Voice)	N/A	Standard	LR=1e-5, steps=4000	32.40
2	Pretrained Large V3 on UASpeech	Whisper Large V3	None (pretrained)	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	No fine-tuning	127.57
3	Pretrained Large V3 on TORGO (1-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	No fine-tuning	108.65
4	Pretrained Large V3 on TORGO (multi-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (multi-word)	M04, M05, F05	454	No fine-tuning	43.17
5	Fine-tune on UASpeech (longer eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=5000, batch=16	— (overfit)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
6	Fine-tune on UASpeech (shorter eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=1500, eval=100	28.41
7	Exp 6 model → UASpeech 1-word test	Exp 6 model	—	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	Inference only	34.53 (avg)
8	Exp 6 model → TORGO 1-word test	Exp 6 model	—	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	Inference only	65.64 (avg)
9	Fine-tune on TORGO multi-word	Whisper Large V3	TORGO (multi-word)	M01,M02,F01,M03,F03	854	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=1000, eval=100	18.77
10	Exp 9 model → TORGO multi-word test	Exp 9 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	17.69 (avg)

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
11	Fine-tune + TORGO + Voice Clone augment	Whisper Large V3	TORGO + Voice Clone	M01,M02,F01,M03,F03	5,578 (854+4,724)	TORGO (multi-word)	M04. M05. F04	454	LR=1e-5, steps=5000, eval=100	15.36
12	Exp 11 model → TORGO multi-word test	Exp 11 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	15.53 (avg)
13	Fine-tune + TORGO + Speech Synthesis	Whisper Large V3	TORGO + Synthesis	M01,M02,F01,M03,F03	5,978 (854+5,124)	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=5000, eval=100	16.84
14	Exp 13 model → TORGO multi-word test	Exp 13 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	16.98 (avg)

Note. WER = Word Error Rate. Data from [redacted], Figure 5.

Note. WER = Word Error Rate. Train/test speaker groups are non-overlapping in all experiments involving fine-tuning. Exp 5 is excluded from WER reporting due to confirmed overfitting before the first evaluation checkpoint.

4. Results

4.1 Why WER Exceeds 100%: An Explanation

Several experiments (2, 3) yield WER values exceeding 100%. This is mathematically possible because WER = (S + I + D) / N, where N is the number of words in the reference transcription. When a model generates extensive hallucinated output—inserting many spurious words not present in the reference—the count of insertions (I) alone can exceed N, pushing WER above 1.0 (100%). For severely dysarthric single-word utterances, the unmodified Whisper Large V3 model frequently produces multi-word hallucinated transcriptions in response to acoustically ambiguous or highly atypical input, resulting in the observed WER values of 122–163%.

4.2 Baseline: Pretrained Whisper Large V3 (Experiments 2–4)

The unmodified Whisper Large V3 model was evaluated on both datasets to establish baselines. Results are presented in Table 2.

Table 2 *Baseline WER of Unmodified Whisper Large V3 by Dataset and Severity*

Table 2

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Experiment	Dataset	Utterance Type	Severe WER (%)	Manageable WER (%)	Mild WER (%)	Overall WER (%)
2	UASpeech	Single-word	142.57	163.77	110.68	127.57
3	TORGO	Single-word	122.95	114.71	86.24	108.65
4	TORGO	Multi-word	90.95	34.08	5.74	43.17

Note. WER = Word Error Rate. Data from [redacted] Figure 5.

The particularly high WER for manageable speakers in UASpeech (163.77%) reflects hallucination behavior: the model's decoder, receiving acoustically ambiguous input

characteristic of moderate dysarthria, produces spurious word sequences longer than the reference. The substantially lower WER for multi-word utterances (Experiment 4) compared to single-word utterances (Experiments 2–3) demonstrates that linguistic context enables the decoder to partially recover from acoustic ambiguity.

4.3 Fine-Tuning on UASpeech (Experiments 5–8)

Experiment 5 revealed that fine-tuning with eval*steps=1000 allowed overfitting before the first evaluation point—training loss reached zero before a model checkpoint was saved with a valid WER. Experiment 6 reduced max*steps to 1,500 and eval_steps to 100, preventing premature convergence and achieving WER=28.41% on the UASpeech test set. The Experiment 6 model was subsequently tested on both UASpeech (Experiment 7) and TORGO single-word utterances (Experiment 8), showing improvements across all severity levels:

Table 3 *WER of Fine-Tuned UASpeech Model (Experiments 7–8) vs. Baseline*

Table 3

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Exp	Test Dataset	Severe WER (%)	Manageable WER (%)	Mild WER (%)	vs. Baseline Improvement
7	UASpeech (1-word)	81.13	18.34	5.56	43.1% / 88.8% / 95.0% better
8	TORGO (1-word)	76.53	70.53	49.87	37.8% / 38.5% / 42.2% better

Note. WER = Word Error Rate. Data from [redacted], Figure 5.

4.4 Fine-Tuning on TORGO Multi-Word Utterances (Experiments 9–10)

Experiment 9 fine-tuned Whisper Large V3 exclusively on TORGO imperative sentences (854 files, 5 training patients, 3 different test patients). Despite the small training set, WER reached

18.77%—a substantial reduction from the 43.17% baseline. Experiment 10 confirmed: WER of 33.04% (severe), 17.52% (manageable), and 2.50% (mild) compared to baseline values of 90.95%, 34.08%, and 5.74%, representing 63.7%, 48.5%, and 56.4% improvements respectively.

4.5 Voice-Cloning Augmentation (Experiments 11–12)

Experiment 11 augmented the 854-sample real training corpus with 4,724 voice-cloned samples (synthetic:real ratio ≈ 5.5:1), expanding training to ~5,578 samples. Training ran for 5,000 steps (14–17 epochs at effective batch size 16). WER on the validation set reached 15.36%—an improvement of 1.41 percentage points over Experiment 9. The augmentation did not substantially alter the severity balance of the training set, as voice cloning was applied uniformly across training-set patient groups.

Experiment 12 tested this model on TORGO multi-word utterances:

Table 4 *Voice-Clone-Augmented Model WER (Experiment 12) vs. Pretrained Baseline*

Table 4

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Severity	Baseline WER (%)	Fine-Tuned + Clone WER (%)	WER Reduction (%)
Severe	90.95	26.89	70.4%
Manageable	34.08	17.04	50.0%
Mild	5.74	2.66	53.7%

Note. WER = Word Error Rate. Data from [redacted]), Figure 5.

4.6 Speech-Synthesis Augmentation (Experiments 13–14)

Experiment 13 replaced voice-cloned data with speech-synthesized samples (~5,124 files; synthetic:real ratio ≈ 6:1). WER reached 16.84% on the validation set—1.48 percentage points higher than the voice-clone model. Experiment 14 confirmed:

Table 5 *Speech-Synthesis-Augmented Model WER (Experiment 14) vs. Voice-Clone Model (Exp 12)*

Table 5

Training Metrics Across Steps (WER, Training Loss, Validation Loss)

Severity	Voice-Clone WER (%)	Speech-Synth WER (%)	Difference
Severe	26.89	29.44	+2.55 (worse)
Manageable	17.04	18.49	+1.45 (worse)
Mild	2.66	3.00	+0.34 (worse)

Note. WER = Word Error Rate. Data from [redacted] Figure 5.

Speech synthesis produced slightly higher WER than voice cloning despite eliminating hallucination artifacts, suggesting that fine-tuning F5-TTS on the full TORGO corpus introduces averaging effects across speakers that reduce the acoustic individuality of the generated samples.

4.7 Training Dynamics: WER vs. Validation Loss Divergence

The detailed training log for Experiment 11 (Figure 5, Appendix 18) illustrates a key finding: validation loss reached its minimum of 0.4121 at step 2,250, while WER continued improving to 15.36% at step 5,000. Had training been stopped at minimum validation loss, WER would have been approximately 17.05%—1.69 percentage points higher than the final model. This empirically justifies WER-based early stopping over validation-loss-based stopping, consistent with Liu et al. (2024).

5. Discussion

5.1 Restatement of Contributions

This study demonstrates that fine-tuning large-scale ASR models can significantly improve recognition accuracy for dysarthric speech, with voice-cloning-based augmentation outperforming speech synthesis across all severity levels.

More importantly, this work provides a systematic empirical evaluation of how state-of-the-art ASR models can be adapted to a highly underserved and clinically relevant domain. The study offers new insights into (1) the effectiveness of synthetic data augmentation strategies, (2) the relationship between dysarthria severity and ASR performance, and (3) the importance of WER-based training optimization in low-resource scenarios.

Together, these contributions extend existing knowledge by demonstrating not only that adaptation is possible, but how it can be practically achieved and deployed in real-world assistive applications.

5.2 Why Severe Dysarthria Produces Higher WER

The consistent severity-WER relationship across all experiments reflects the underlying acoustic properties of dysarthric speech. Severe dysarthria involves pronounced breakdown of articulatory precision: phonemes are produced with reduced acoustic distinctiveness, co-articulation patterns become irregular, and prosodic cues (stress, rhythm, intonation) that support ASR decoding are largely absent (Young & Mihailidis, 2010). The Whisper encoder, trained on typical speech spectrograms, produces degraded latent representations for severely atypical input, and the decoder—which relies on both acoustic evidence and language model priors—defaults to higher-probability word sequences that may not match the reference (hallucinations). This explains the observed WER >100% for severe speakers under the unmodified baseline (Experiments 2–3).

After fine-tuning with dysarthric training data, the encoder learns to map atypical spectrograms to more appropriate latent representations, and the decoder learns dysarthric-specific acoustic-phonetic correspondences. However, the improvement is largest for severe

speakers (70% WER reduction) precisely because the baseline is highest—leaving the most room for improvement—rather than because severe speech is inherently easier to model after fine-tuning.

5.3 Limitations of WER as a Sole Metric

WER does not distinguish between substitution, insertion, and deletion error types, which carry different clinical implications. Substitution errors (incorrect word recognized) are typically less disruptive than deletion errors (words missed entirely) for communication-aid applications. Future work should report error-type breakdowns per severity level.

The absence of confidence intervals or cross-run variance in this study is a limitation. Each experiment was conducted as a single training run due to computational cost constraints (each run required 1.5–6.5 hours on a dedicated GPU server). Variance estimation through repeated runs or cross-validation over speaker partitions should be incorporated in future work to strengthen statistical claims.

5.4 Comparison to Prior Dysarthric ASR Systems

Schu et al. (2022) reported consistent WER improvement when using ASR systems specifically adapted for UASpeech and TORGO, establishing that domain-specific adaptation is necessary. Rathod et al. (2023) demonstrated that transfer learning with Whisper improves dysarthric transcription accuracy, consistent with the findings of this study. The WER values achieved here (15.36–18.77% for multi-word TORGO utterances with fine-tuning) compare favorably to the baseline values reported in Schu et al. (2022) for unadapted systems, though direct numerical comparison is constrained by differences in dataset partitioning and evaluation protocols.

5.5 Voice Cloning vs. Speech Synthesis

Voice cloning outperformed speech synthesis in all severity categories (Experiments 12 vs. 14). This is likely because voice cloning preserves the speaker-specific acoustic fingerprint of individual dysarthric patients, including their idiosyncratic articulatory patterns, whereas speech synthesis averages over the full training population, producing samples that are acoustically less representative of any specific speaker's impairment profile. The manual quality-control step (removing 185 hallucinated voice-cloned files) was essential to the quality of the augmented dataset.

5.6 Contextual Information and Single vs. Multi-Word Utterances

The dramatic WER difference between single-word (Experiments 2–3: WER >100%) and multi-word utterances (Experiment 4: WER=43.17% baseline; Experiment 11: WER=15.36% fine-tuned) reflects the language model's ability to leverage sequential context. In multi-word sequences, the Whisper decoder uses probability distributions over preceding tokens to constrain candidate words, partially compensating for degraded acoustic features. Single-word utterances provide no such context, forcing the decoder to rely entirely on a single degraded acoustic representation—leading to hallucination in severely impaired cases.

5.7 Practical Application: ClearVoiceAI

The fine-tuned model was deployed in ClearVoiceAI, a web application (clearvoiceai.com) enabling dysarthric users to record speech or upload audio files for real-time transcription. The application stack comprises a JavaScript/CSS frontend, a Python/FastAPI backend, and the fine-tuned Whisper model hosted on AWS. The voice-clone-augmented model (Experiment 11) serves as the default ASR engine. Practical deployment revealed that environmental noise and spontaneous (non-scripted) dysarthric speech present additional challenges beyond those captured in the controlled-environment TORGO and UASpeech recordings.

5.8 Limitations

- Single training runs per configuration; no cross-run variance estimates reported
- No error-type breakdown (substitutions, insertions, deletions) per severity
- Evaluation limited to controlled-environment recordings (TORGO, UASpeech); real-world acoustic conditions not tested
- Training data limited to ~5,600 samples; larger corpora would likely improve generalization
- Speaker-specific fine-tuning (personalizing models for individual patients) was not explored

5.9 Future Directions

- Longitudinal WER monitoring to track dysarthria progression as a clinical tool
- Speaker-adaptive fine-tuning: personalizing models for individual patients in real-time
- Extension to other speech disorders (dysphagia, apraxia, aphasia)
- Cross-linguistic evaluation (dysarthric speech in languages other than English)
- Real-world noise robustness evaluation
- Edge deployment for on-device inference without cloud dependency

6. Conclusion

This study empirically evaluated fine-tuning strategies for adapting Whisper Large V3 to dysarthric speech, using the publicly available UASpeech and TORGO datasets with strictly non-overlapping speaker partitions between training and test sets. Across 14 systematic experiments, voice-cloning augmentation produced the best-performing model (WER=15.36% on TORGO multi-word utterances), with severity-specific reductions of approximately 70% (severe), 50% (moderate), and 54% (mild) relative to the unmodified pretrained baseline.

The primary contributions of this work lie in the systematic application and evaluation of state-of-the-art ASR adaptation techniques to the underserved domain of dysarthric speech recognition. Through 14 controlled experiments, the study provides new empirical insights into the effectiveness of fine-tuning strategies, the comparative impact of voice cloning versus speech synthesis, and the influence of severity-specific acoustic variability on model performance.

In addition to these experimental findings, the deployment of the fine-tuned model in the ClearVoiceAI system demonstrates the practical feasibility of translating research advances into real-world assistive technology. These contributions collectively strengthen both the scientific understanding and applied capabilities of dysarthric speech recognition systems.

7. Code and Data Availability

The fine-tuned Whisper models are publicly available on Hugging Face. Training code is maintained on GitHub. Training metrics, loss curves, and WER logs are available on WandB.AI.

- **Code repository:** GitHub — Fine-tuning and inference scripts
- **Model repository:** Hugging Face — Fine-tuned Whisper Large V3 models
- **Training metrics:** WandB.AI — Training loss, validation loss, WER logs across all experiments
- **Application:** ClearVoiceAI — Deployed ASR web application

References

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2024). F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*. https://arxiv.org/abs/2410.06885

Farhadipour, A., & Veisi, H. (2023). Gammatonegram representation for end-to-end dysarthric speech processing tasks: Speech recognition, speaker identification, and intelligibility assessment. *arXiv preprint arXiv:2307.03296*. https://doi.org/10.48550/arxiv.2307.03296

Gandhi, S. (2022, November 3). *Fine-tune Whisper for multilingual ASR with Transformers*. Hugging Face. https://huggingface.co/blog/fine-tune-whisper

Kim, H., Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T., Watkin, K., Frame, S., Sharma, H. V., & Zhou, X. (2023). *UASpeech.* IEEE Dataport. https://dx.doi.org/10.21227/f9tc-ab45

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019). Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*. https://arxiv.org/abs/1912.11370

Lettergram. (2019). *Sentence classification — imperatives dataset.* GitHub. https://github.com/lettergram/sentence-classification

Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. *Journal of Audio, Speech, and Music Processing, 2024*(29). https://doi.org/10.1186/s13636-024-00349-3

Pennington, L., Roelant, E., Thompson, V., Robson, S., Steen, N., & Miller, N. (2013). Intensive dysarthria therapy for younger children with cerebral palsy. *Developmental Medicine & Child Neurology*, *55*(5), 464–471. https://doi.org/10.1111/dmcn.12098

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. https://arxiv.org/abs/2212.04356

Rathod, B., et al. (2023). Transfer learning using Whisper for dysarthric automatic speech recognition. In *Proceedings of the International Conference on Speech Technologies* (pp. 419–431). Springer. https://doi.org/10.1007/978-3-031-48309-7_46

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*(3), 523–541. https://doi.org/10.1007/s10579-011-9145-0

Schölderle, T., Haas, E., & Ziegler, W. (2020). Dysarthria syndromes in children with cerebral palsy. *Developmental Medicine & Child Neurology*, *63*(4), 444–449.

https://doi.org/10.1111/dmcn.14679

Schu, G., Janbakhshi, P., & Kodrasi, I. (2022). On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. *arXiv preprint arXiv:2211.08833*. https://arxiv.org/abs/2211.08833

Seagraves, A. (2022, December 19). *3 best open-source ASR models compared: Whisper, wav2vec 2.0, Kaldi.* Deepgram. https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., & Shih, M.-H. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *Healthcare, 10*(10), 1956. https://doi.org/10.3390/healthcare10101956

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis, 11*(1–2), 4–15. https://doi.org/10.3109/17482960802379004

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. https://doi.org/10.1080/10400435.2010.483646

U.S. Patent Documents Cited by Examiner

Burkhardt. (2016, April). *Speech synthesis system* (U.S. Patent Publication No. 2016/0104477 A1). U.S. Patent and Trademark Office. [CPC: G10L 13/02]

Burns. (2020, September). *Speech recognition for disordered speech* (U.S. Patent Publication No. 2020/0279549 A1). U.S. Patent and Trademark Office. [CPC: G10L 21/003]

Chang. (2022, September). *Acoustic model adaptation* (U.S. Patent Publication No. 2022/0301563 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/24]

Ingel. (2025, January). *Natural language processing system* (U.S. Patent Publication No. 2025/0006182 A1). U.S. Patent and Trademark Office. [CPC: G06F 40/30]

Kanevsky. (2014, July). *Personalized speech recognition* (U.S. Patent Publication No. 2014/0214426 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/08]

Koul. (2023, July). *Communication assistance system* (U.S. Patent No. 11,699,360 B2). U.S. Patent and Trademark Office. [CPC: H04M 3/42391]

Krishna. (2023, May). *Automatic speech recognition with domain adaptation* (U.S. Patent Publication No. 2023/0139394 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/24]

Li. (2024, October). *Neural network-based speech processing* (U.S. Patent Publication No. 2024/0347064 A1). U.S. Patent and Trademark Office. [CPC: G06N 3/045]

Lin. (2020, October). *End-to-end speech recognition* (U.S. Patent Publication No. 2020/0312302 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/063]

Lin. (2021, July). *Speech enhancement using deep learning* (U.S. Patent Publication No. 2021/0225384 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/66]

McNair. (2023, September). *Speaker-adaptive speech processing* (U.S. Patent Publication No. 2023/0290353 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/66]

McNulty. (2024, October). *Speech recognition for atypical speakers* (U.S. Patent Publication No. 2024/0361827 A1). U.S. Patent and Trademark Office. [CPC: G10L 25/63]

Phillips. (2011, March). *Dysarthric speech recognition system* (U.S. Patent Publication No. 2011/0054896 A1). U.S. Patent and Trademark Office. [CPC: G10L 15/30]

Sharma. (2025, March). *Speech synthesis and recognition* (U.S. Patent Publication No. 2025/0104689 A1). U.S. Patent and Trademark Office. [CPC: G10L 21/10]

Wang. (2025, March). *Sequence-to-sequence speech model* (U.S. Patent No. 12,249,324 B1). U.S. Patent and Trademark Office. [CPC: G10L 25/27]

Appendix

Patent Figures from U.S. Patent No. [redacted] — [redacted] (2025).

[redacted]

U.S. Patent Dec. 2, 2025 Sheet 1 of 15 US 12,488,786 B1

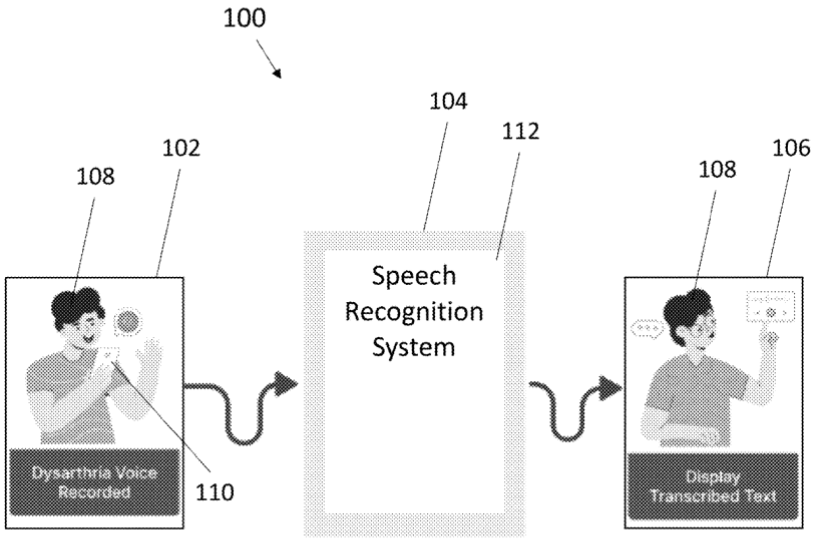


FIG. 1

Overview of the speech recognition system for dysarthric speech, illustrating the end-to-end flow from voice recording to transcribed text display.

Source: Patent FIG. 1, Patent FIG. 1. U.S. Patent No. [redacted]

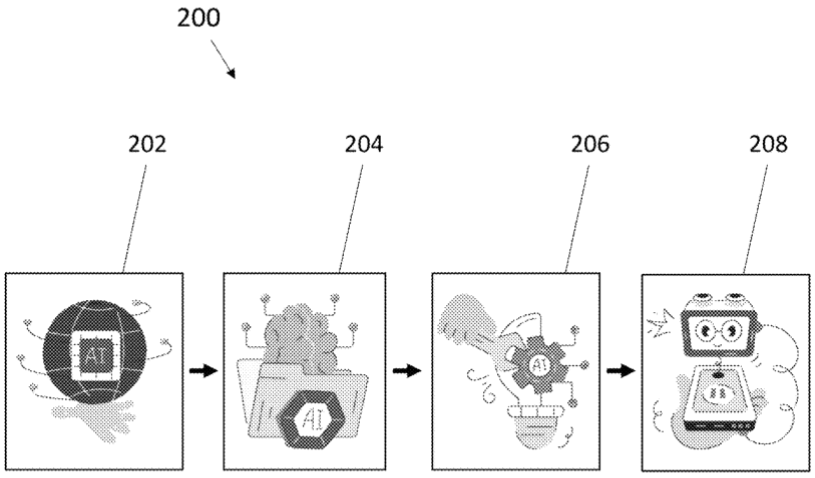


FIG. 2

Four-stage processing pipeline depicting AI-powered speech recognition: input acquisition, feature extraction, model inference, and output generation.

Source: Patent FIG. 2, Patent FIG. 2. U.S. Patent No. [redacted]

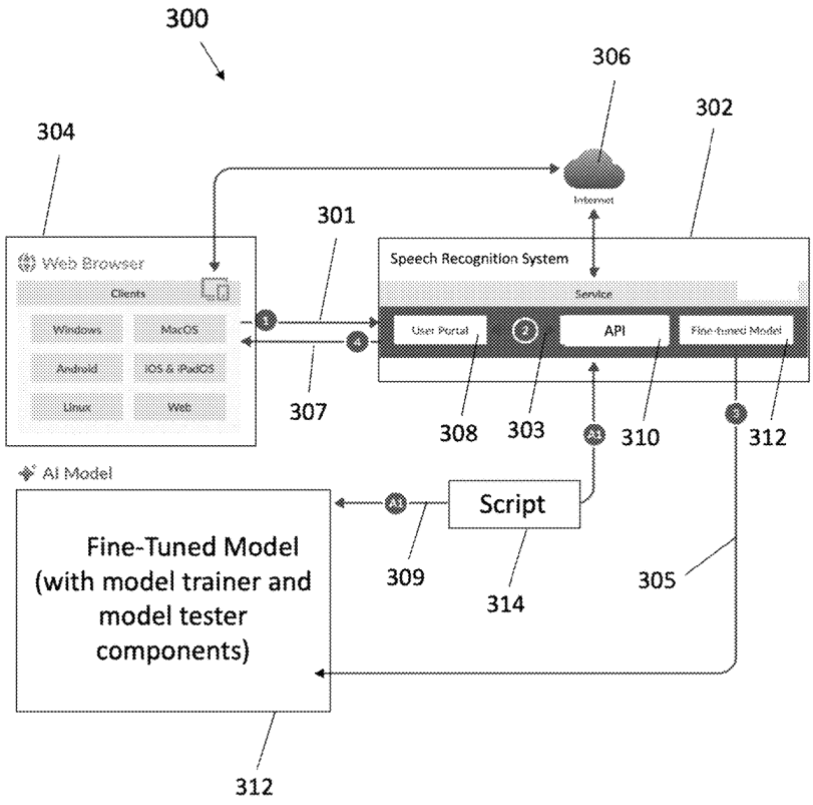


FIG. 3

System architecture showing web browser clients, the Speech Recognition System service layer (User Portal, API, Fine-tuned Model), and the model training infrastructure with trainer and

tester components.

Source: Patent FIG. 3, Patent FIG. 3. U.S. Patent No. [redacted]

400

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./voice-clone-large-finetune-final",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=2,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=5000,
    gradient_checkpointing=True,
    fp16=True,
    eval_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=250,
    eval_steps=250,
    logging_steps=25,
    report_to=["wandb"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=True,
    save_total_limit=2
)
```

FIG. 4

Training configuration parameters for Seq2Seq model fine-tuning, including learning rate, batch size, gradient accumulation, warmup steps, and WER-based evaluation settings.

Source: Patent FIG. 4, Patent FIG. 4. U.S. Patent No. [redacted]

500

Training Loss	Epoch	Step	Validation Loss	Wer
0.1607	0.8460	250	0.5163	25.9413
0.0598	1.6920	500	0.4849	24.8444
0.0257	2.5381	750	0.4450	30.4180
0.0141	3.3841	1000	0.4369	19.3003
0.0029	4.2301	1250	0.4267	16.0095
0.0015	5.0761	1500	0.4209	18.4109
0.0063	5.9222	1750	0.4259	19.3300
0.0016	6.7682	2000	0.4341	17.7587
0.0009	7.6142	2250	0.4121	17.0471
0.0013	8.4602	2500	0.4199	16.3653
0.0009	9.3063	2750	0.4233	16.5135
0.001	10.1523	3000	0.4237	16.0688
0.0019	10.9983	3250	0.4230	16.4542
0.0014	11.8443	3500	0.4292	15.8316
0.0007	12.6904	3750	0.4291	15.8316
0.0005	13.5364	4000	0.4321	15.3869
0.0009	14.3824	4250	0.4334	15.2980
0.001	15.2284	4500	0.4344	15.2980
0.0	16.0745	4750	0.4372	15.3572
0.0	16.9205	5000	0.4377	15.3572

502

504

FIG. 5

Full training metrics log across 5,000 optimization steps, showing training loss, epoch, validation loss, and Word Error Rate (WER) at each checkpoint.

Source: Patent FIG. 5, Patent FIG. 5. U.S. Patent No. [redacted]

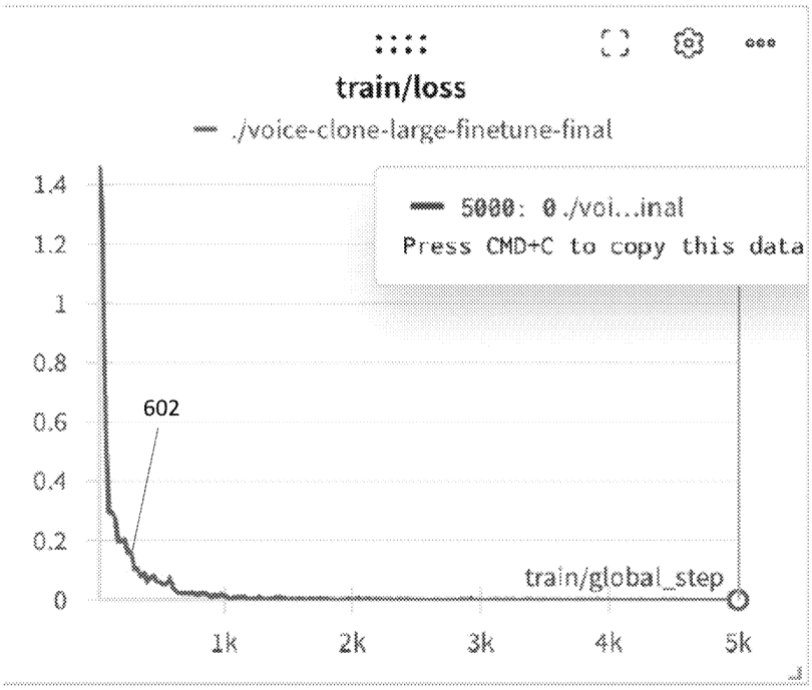


FIG. 6A

Training loss convergence curve across global training steps, demonstrating rapid decrease from ~1.4 to near zero by step 5,000.

Source: Patent FIG. 6A, Patent FIG. 6A. U.S. Patent No. [redacted]

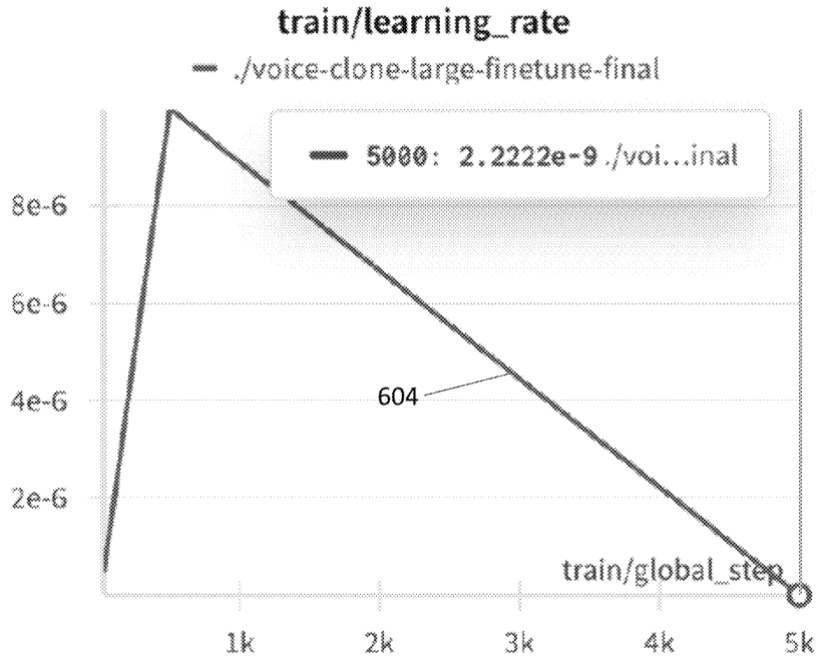


FIG. 6B

Learning rate schedule over the course of fine-tuning, showing a linear warmup phase followed by linear decay from ~1e-5 to near zero.

Source: Patent FIG. 6B, Patent FIG. 6B. U.S. Patent No. [redacted]

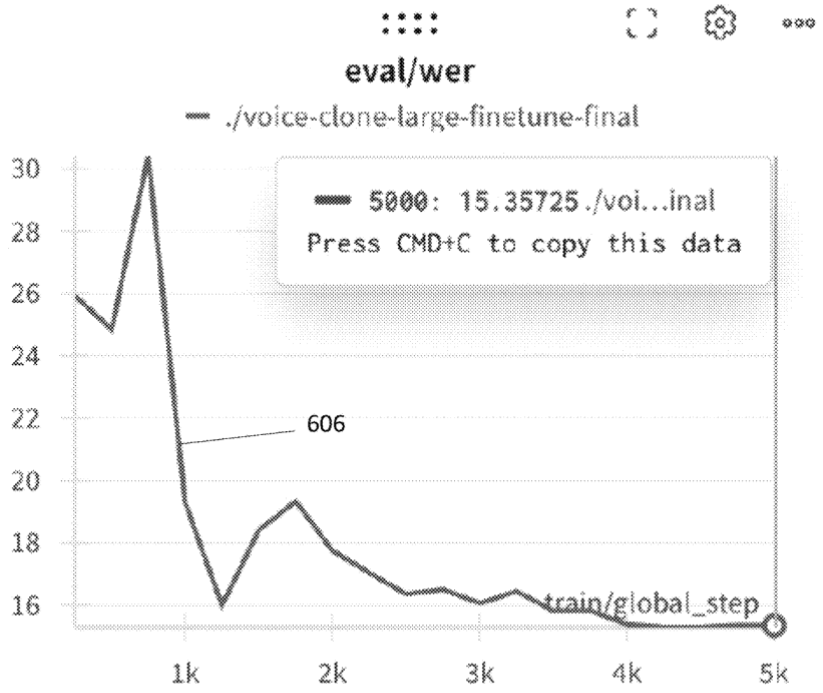


FIG. 6C

Word Error Rate (WER) trajectory across training steps, illustrating non-monotonic convergence with a final WER of approximately 15.36% at step 5,000.

Source: Patent FIG. 6C, Patent FIG. 6C. U.S. Patent No [redacted]

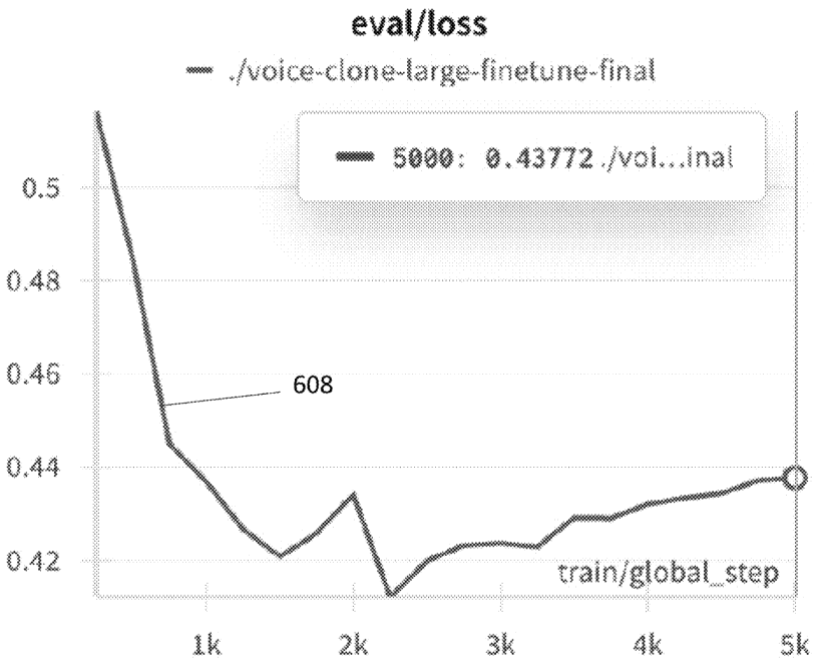


FIG. 6D

Evaluation loss across training steps, reaching its minimum of ~0.4121 at step 2,250 and diverging slightly thereafter, illustrating the WER vs. validation loss discrepancy discussed in

Section 5.3.

Source: Patent FIG. 6D, Patent FIG. 6D. U.S. Patent No [redacted]

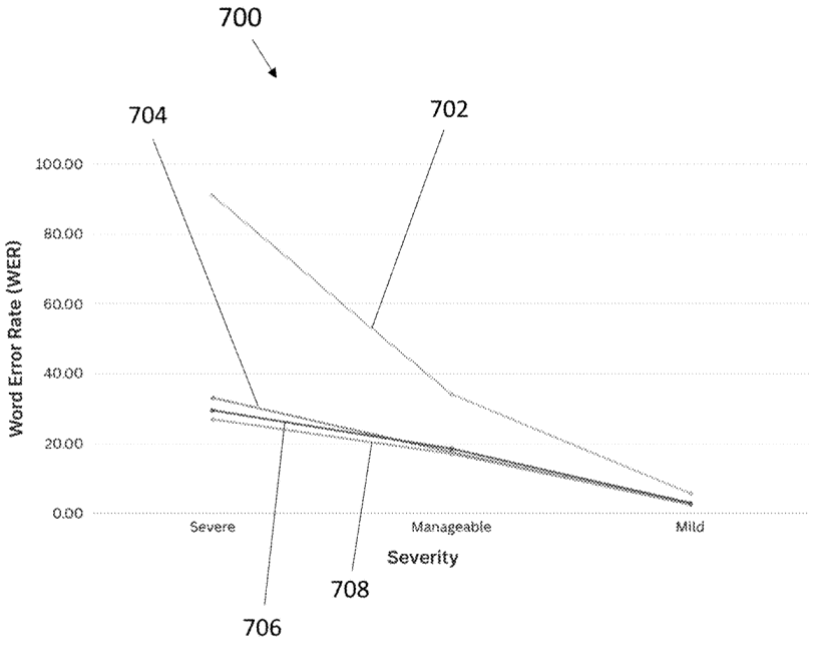


FIG. 7

WER comparison across speech impairment severity levels (Severe, Manageable, Mild) for baseline versus fine-tuned ASR models, demonstrating the greatest improvement for severely

impaired speakers.

Source: Patent FIG. 7, Patent FIG. 7. U.S. Patent No. [redacted]

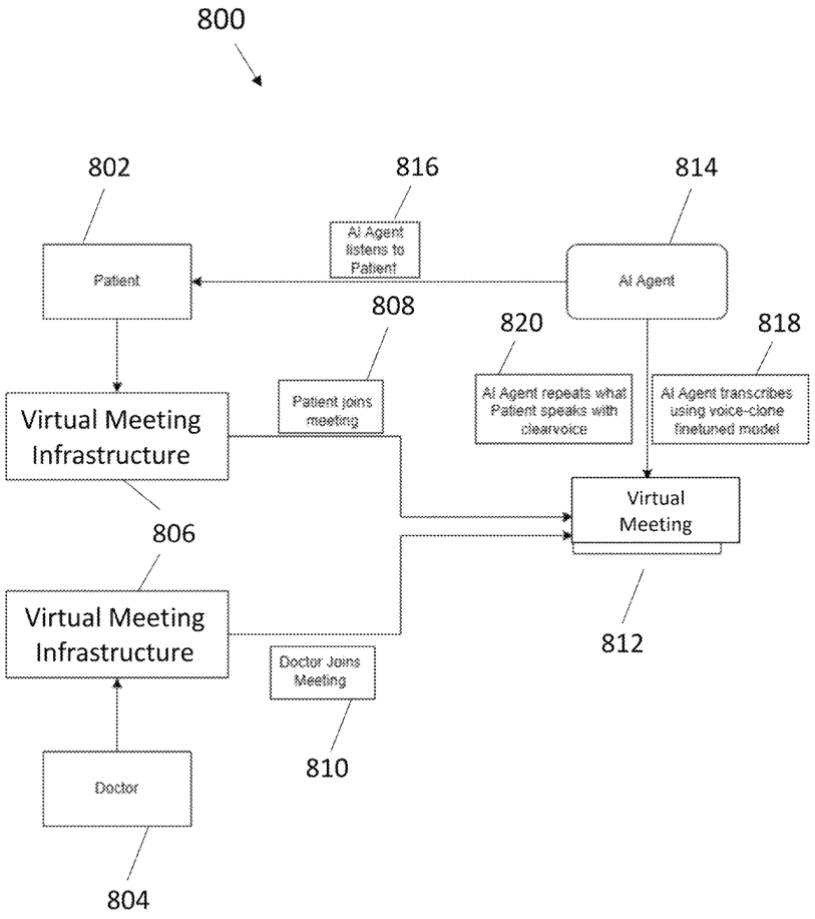


FIG. 8

Virtual meeting AI agent architecture enabling dysarthric patients to participate in telemedicine and virtual consultations via AI-powered transcription and clear-voice

re-vocalization.

Source: Patent FIG. 8, Patent FIG. 8. U.S. Patent No[redacted]

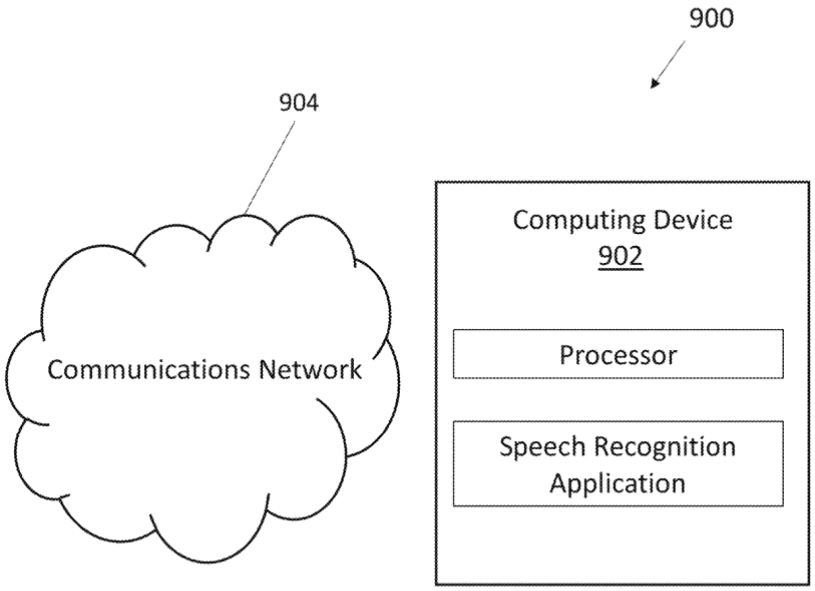


FIG. 9

Computing system network architecture for speech recognition and transcription, showing the computing device connected via communications network.

Source: Patent FIG. 9, Patent FIG. 9. U.S. Patent No. [redacted]

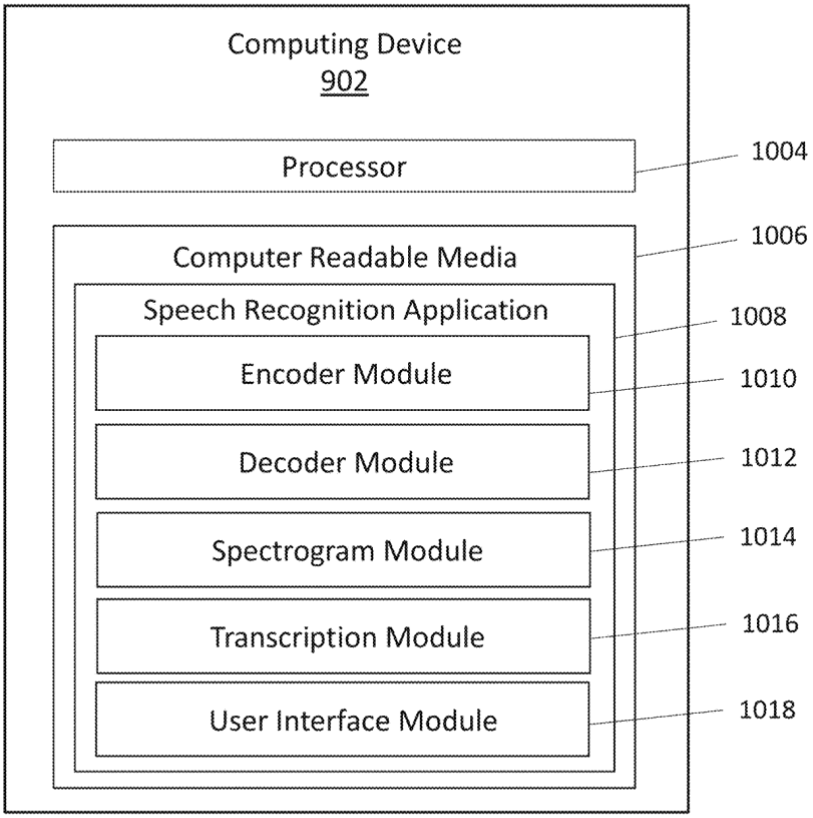


FIG. 10

Computing device (902) architecture with processor and speech recognition application comprising Encoder Module, Decoder Module, Spectrogram Module, Transcription Module,

and User Interface Module.

Source: Patent FIG. 10, Patent FIG. 10. U.S. Patent No[redacted]

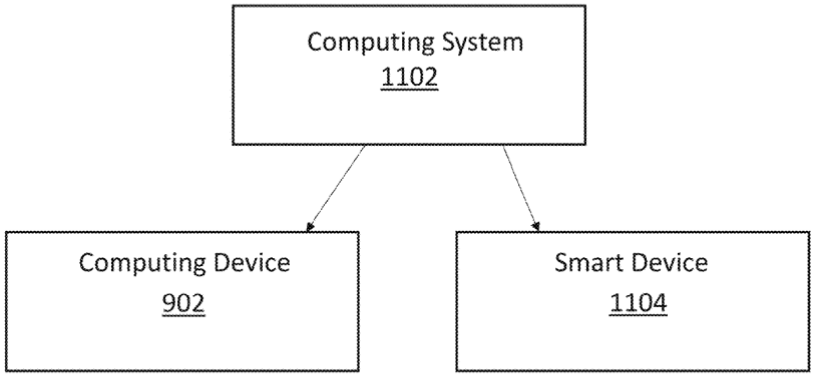


FIG. 11

System hierarchy showing the computing system (1102) integrating a Computing Device (902) and a Smart Device (1104) for voice-controlled IoT applications.

Source: Patent FIG. 11, Patent FIG. 11. U.S. Patent No [redacted]

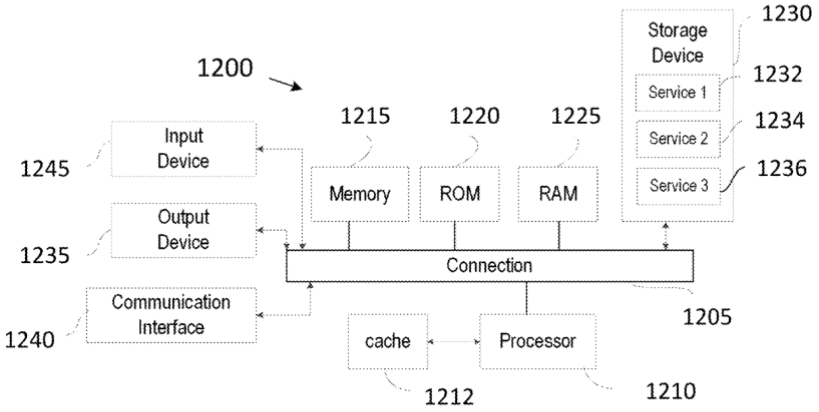


FIG. 12

Detailed computing system block diagram illustrating processor, memory hierarchy (Memory, ROM, RAM), storage services, I/O devices, and communication interface components.

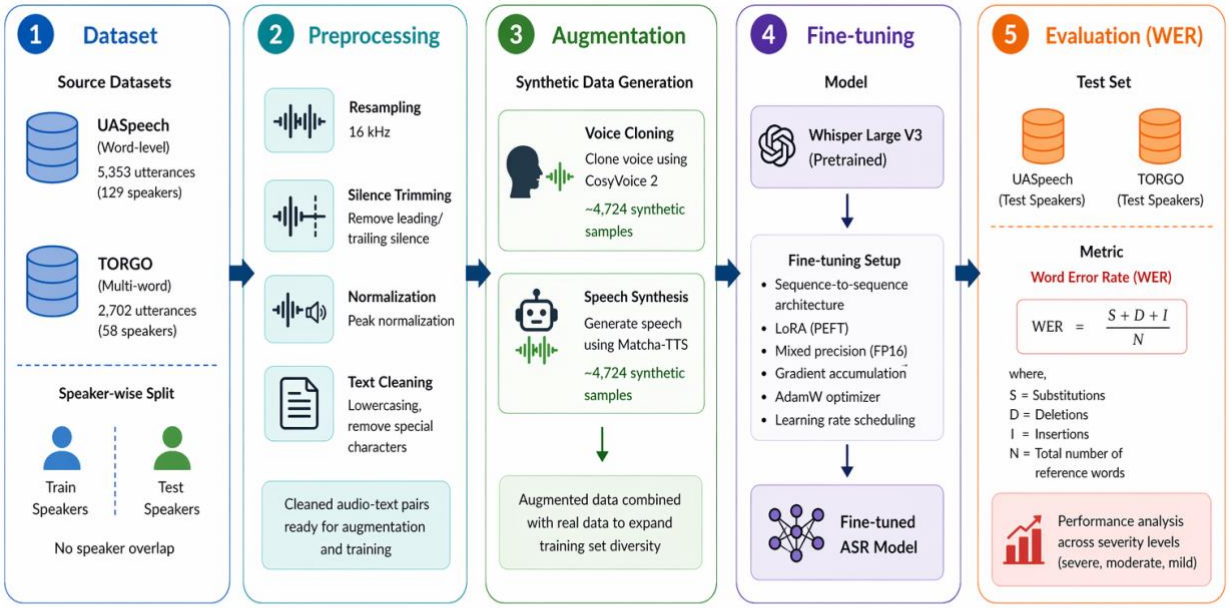


FIG. 13

Overall analytical methodology pipeline illustrating the end-to-end process: dataset collection (UASpeech and TORGO), audio preprocessing (resampling, silence trimming, normalization), synthetic data augmentation (voice cloning and speech synthesis), fine-tuning of Whisper Large V3, and evaluation using Word Error Rate (WER) across dysarthria severity levels.

TRACKED CHANGES MANUSCRIPT

TRACKED CHANGES LEGEND: ~~Deleted text~~ Added text

FINE-TUNED SPEECH RECOGNITION FOR DYSARTHRIC SPEECH

SPEECH RECOGNITION FOR ASSISTING PATIENTS WITH SPEECH DIFFICULTIES

[redacted]

[redacted]

Author Note

Correspondence concerning this article should be addressed t[redacted]

Competing Interests: The authors declare that the methods and systems described in this article are the subject of U.S. Patent [redacted]

Abstract

Automated speech recognition (ASR) systems based on large pretrained models achieve near-human accuracy for typical speakers, yet consistently fail for individuals with dysarthria—a motor speech disorder affecting articulation, phonation, and prosody. This paper presents an empirical evaluation of fine-tuning strategies for adapting OpenAI Whisper Large V3, a state-of-the-art sequence-to-sequence ASR model, to dysarthric speech drawn from two established public datasets: UASpeech and TORGO. The study also examines the applied contribution of voice-cloning and speech-synthesis-based data augmentation as a practical approach to the chronic low-resource challenge in dysarthric ASR. Fourteen systematic experiments were conducted, varying training data composition, augmentation strategy, and hyperparameter configuration. Speaker groups were strictly separated between training and test partitions to ensure uncontaminated evaluation. The fine-tuned Whisper Large V3 model augmented with approximately 4,724 voice-cloned samples achieved a Word Error Rate (WER) of 15.36% on multi-word TORGO utterances—representing reductions of approximately 70% for severe, 50% for moderate, and 54% for mild dysarthric speech compared to the unmodified pretrained baseline. A deployed ASR web application, ClearVoiceAI, demonstrates the practical utility of the approach. These findings contribute an empirical foundation for applying established fine-tuning and data augmentation techniques to the underserved domain of dysarthric speech recognition.

Keywords: dysarthria, automatic speech recognition, Whisper, fine-tuning, voice cloning, word error rate, UASpeech, TORGO, data augmentation, assistive technology

1. Introduction

~~1.1 Background and Motivation~~ 1.1 Dysarthria and Automated Speech Recognition Systems

Dysarthria is a motor speech disorder ~~caused by damage to the neurological mechanisms governing articulation, respiration, and phonation~~ resulting from neurological impairment that affects the coordination, strength, and control of the muscles involved in speech production (Tomik & Guiloff, 2010). ~~It is defined by reduced strength, speed, range, tone, or coordination of the muscles involved in speech production (Pennington et al., 2013). The disorder is strongly associated with neurological conditions including cerebral palsy (CP), which affects approximately 40% of its patients with dysarthria, and amyotrophic lateral sclerosis (ALS), which affects up to 80%~~ Dysarthria is strongly associated with neurological conditions such as cerebral palsy (CP), affecting approximately 40% of individuals with CP, and amyotrophic lateral sclerosis (ALS), where prevalence can reach up to 80% (Shih et al., 2022). Additional ~~conditions linked to~~ associated conditions include stroke, Parkinson's disease, multiple sclerosis, and traumatic brain injury (Schölderle et al., 2020).

The ~~effects of dysarthria on speech~~acoustic manifestations of dysarthria include irregular articulation, slurred ~~phonemes,~~ or imprecise phoneme production, atypical prosody, reduced ~~intelligibility, and unpredictable fluctuations in acoustic quality~~vocal stability, and inconsistent speech patterns (Young & Mihailidis, 2010). These characteristics vary significantly across individuals and severity levels, creating high variability in speech signals. For affected individuals, these characteristics create significant communication barriers in daily life, ~~including in technology-mediated contexts.~~impacting social participation, personal identity, and overall quality of life (Vogel et al., 2026; Page & Yorkston, 2022). Individuals with dysarthria often experience reduced confidence in communication, social isolation, and stigma associated with impaired speech.

~~[Section 1.2 Problem Statement — deleted and content merged into Section 1.1]. Contemporary commercial ASR systems—including virtual assistants and speech-to-text tools—are trained predominantly on speech from neurotypical speakers. When applied to dysarthric speech, these systems exhibit severe accuracy degradation. Young and Mihailidis (2010) documented consistent failure of standard ASR systems for moderate-to-severely dysarthric speakers. The baseline evaluation in this study confirms this: the unmodified Whisper Large V3 model—one of the most capable publicly available ASR models—achieved WER values exceeding 100% on dysarthric speech from both the UASpeech and TORGO datasets (see Section 4).~~

~~This level of error renders standard systems unusable as communication aids for affected patients.~~

Automatic speech recognition (ASR) systems have achieved near-human performance for neurotypical speech through large-scale pretraining on diverse datasets. However, these systems exhibit severe performance degradation when applied to dysarthric speech, due to a fundamental domain mismatch between the training data and the target input conditions. Prior studies have consistently demonstrated that standard ASR systems fail to generalize effectively to dysarthric speech, particularly for moderate-to-severe cases (Young & Mihailidis, 2010; Schu et al., 2022).

This limitation is particularly evident in technology-mediated contexts, such as voice assistants, telemedicine systems, and speech-to-text interfaces, where reliance on visual and contextual cues is reduced. These challenges highlight the need for targeted adaptation strategies to improve accessibility for individuals with speech impairments.

1.2 Adapting Large-Scale ASR Models to Dysarthric Speech: Fine-Tuning and Synthetic Data Strategies

Recent advances in large-scale ASR models, such as Whisper (Radford et al., 2022), have demonstrated robust performance across a wide range of acoustic conditions due to training on extensive multilingual and weakly supervised datasets. However, their effectiveness in low-resource and atypical speech domains, including dysarthria, remains limited without domain-specific adaptation.

Fine-tuning has emerged as a primary approach for adapting pretrained ASR models to specialized domains. By updating model parameters using domain-specific datasets, fine-tuning enables the model to learn acoustic and phonetic patterns unique to dysarthric speech. Prior work has shown that transfer learning can significantly improve transcription accuracy for dysarthric speakers (Rathod et al., 2023). However, the success of fine-tuning is constrained by the limited availability of labeled dysarthric speech data, increasing the risk of overfitting (Liu et al., 2024).

To address this challenge, data augmentation techniques such as voice cloning and speech synthesis have been explored. Voice cloning generates synthetic samples that preserve speaker-specific acoustic characteristics, while speech synthesis produces generalized dysarthric-like speech patterns without requiring per-speaker reference audio. These approaches provide scalable methods for expanding training datasets and improving model robustness in low-resource scenarios.

~~1.3 Scope and Contribution of This Work~~ 1.3 Scope and Aims

~~This study is an empirical evaluation of how well fine-tuning and data augmentation strategies—specifically voice cloning and speech synthesis—can adapt Whisper Large V3 to dysarthric speech. The methods employed (transfer learning, fine-tuning, TTS-based augmentation) are not presented as novel algorithmic contributions; rather, the novel contribution of this work lies in their systematic applied evaluation in the domain of dysarthric ASR, using publicly available dysarthric speech corpora with severity-stratified assessment.~~

This study evaluates the effectiveness of fine-tuning and data augmentation strategies for adapting Whisper Large V3 to dysarthric speech. The primary objective is to assess how voice cloning and speech synthesis can mitigate data scarcity and improve transcription accuracy across varying levels of dysarthria severity.

~~Specifically, this paper contributes:~~

~~1. A systematic comparison of 14 fine-tuning configurations for adapting Whisper Large V3 to dysarthric speech, with full experimental detail including dataset partitions and speaker assignments.~~

~~2. An applied evaluation of voice cloning (F5-TTS) and speech synthesis as low-cost data augmentation strategies for the low-resource dysarthric speech domain.~~

~~3. Severity-stratified WER analysis across severe, moderate, and mild dysarthric speech from UASpeech and TORGO.~~

~~4. A deployed ASR application (ClearVoiceAI) demonstrating real-world applicability.~~

~~5. An empirical justification for using WER stabilization rather than validation loss as the early stopping criterion during fine-tuning.~~

This work contributes by providing a systematic empirical evaluation of state-of-the-art ASR adaptation techniques applied to dysarthric speech datasets. It offers new insights into the relative effectiveness of augmentation strategies, the impact of severity-specific acoustic variability, and the practical challenges of deploying ASR systems in real-world assistive contexts.

1.4 Research Hypothesis

It is hypothesized that fine-tuning Whisper Large V3 on dysarthric speech datasets, augmented with voice-cloned and speech-synthesized samples, will ~~substantially~~significantly reduce ~~WER~~Word Error Rate (WER) compared to the unmodified pretrained model across all ~~severity levels.~~levels of dysarthria severity.

1.5 Paper Organization

~~[Section 1.5 entirely removed in revised version] Section 2 reviews the relevant literature on dysarthric ASR, Whisper, and data augmentation. Section 3 describes the datasets, preprocessing, model architecture, and experimental methodology. Section 4 presents results across all 14 experiments. Section 5 discusses the findings and their implications. Section 6 concludes with limitations and future directions. A code and data availability statement is provided in Section 7.~~

2. Literature Review

2.1 Dysarthria and Its Impact on ASR

The acoustic characteristics of dysarthric speech—including reduced articulatory precision, irregular vocal quality, atypical prosody, and co-articulation breakdown—create a substantial domain mismatch with the training distributions of standard ASR systems (Young & Mihailidis, 2010). Schu et al. (2022) demonstrated that standard ASR systems trained without dysarthric-specific adaptation produce consistently poor results on both the UASpeech and TORGO benchmarks, establishing these datasets as appropriate evaluation grounds for dysarthric ASR research.

2.2 Whisper as a Base Model

Radford et al. (2022) introduced Whisper, a sequence-to-sequence ASR model trained on 680,000 hours of multilingual weakly-supervised web audio, achieving robust performance across diverse acoustic conditions. The Whisper Large V3 variant supports 1.5 billion parameters with multilingual capability. Its architecture—an encoder-decoder transformer with a log-Mel spectrogram front-end—is well-suited to fine-tuning via the Hugging Face transformers library. Liu et al. (2024) systematically evaluated Whisper fine-tuning strategies for low-resource ASR scenarios and identified key hyperparameter sensitivities relevant to small-dataset fine-tuning, which informed the experimental design of this study.

2.3 Fine-Tuning for Dysarthric Speech

Rathod et al. (2023) demonstrated that transfer learning applied to Whisper significantly improves word recognition accuracy for dysarthric speech, establishing fine-tuning as a viable adaptation strategy. The challenge in this domain is that dysarthric speech corpora are inherently small—collecting and annotating such data requires clinical access and significant manual effort—making standard fine-tuning approaches susceptible to overfitting (Liu et al., 2024).

2.4 Voice Cloning and Speech Synthesis as Data Augmentation

To address ~~the data scarcity problem~~data scarcity, this study employs F5-TTS (Chen et al., 2024), a flow-matching-based TTS model ~~capable of voice cloning from~~that can clone voices from short reference audio clips. F5-TTS ~~is used to generate~~generates synthetic dysarthric utterances that preserve the acoustic characteristics of specific impaired speakers—including their articulatory irregularities. An alternative approach, speech synthesis, fine-tunes the base F5-TTS model on the complete TORGO dataset to synthesize novel dysarthric-like utterances

without per-utterance reference audio, avoiding hallucination artifacts common in voice cloning when reference audio quality is low.

~~2.5 Evaluation Metric: Word Error Rate~~ 2.5 Evaluation Metric: Word Error Rate (WER)

Word Error Rate (WER) is the standard evaluation metric for ASR systems. It is computed as:

WER = (S + I + D) / N

where S = substitutions, I = insertions, D = deletions, and N = number of reference words. Importantly, WER can exceed 100% when the number of insertion errors alone surpasses the total number of reference words—a phenomenon observed in this study for severely dysarthric single-word utterances, where the unmodified Whisper model generates extensive hallucinated output (see Section 4.1 and Discussion Section 5.3).

3. Methodology

3.1 Base Model Selection

Three candidate ASR models were evaluated: Kaldi, Meta Wav2Vec 2.0, and OpenAI Whisper (Seagraves, 2022). Based on comparative benchmarking, Whisper's overall WER was 45% lower than Wav2Vec 2.0 and 63% lower than Kaldi across diverse acoustic conditions. OpenAI Whisper Large V3 (Radford et al., 2022) was selected as the base pretrained model for all fine-tuning experiments. Reference: Figure 13.

3.2 Datasets

3.2.1 UASpeech Dataset

The UASpeech dataset (Kim et al., 2023) contains recordings from 15 individuals with dysarthria caused by cerebral palsy and 13 neurotypical controls, comprising approximately 57,000 predominantly single-word utterances. Speakers were classified into severity groups based on speech intelligibility percentages following Farhadipour and Veisi (2023):

- Severe (0–40% intelligibility): Patients F03, M12 — 5,185 test files
- Moderate/Manageable (40–80% intelligibility): Patients M06, M11 — 2,847 test files
- Mild (>80% intelligibility): Patients M08, F05 — 10,711 test files

Training used 9 patients (~38,700 files); testing used 6 different patients (~18,700 files). Speaker groups were strictly non-overlapping between training and test partitions.

3.2.2 TORGO Dataset

The TORGO dataset (Rudzicz et al., 2012) contains recordings from 8 dysarthric speakers (with CP and ALS) and 7 controls. A subset of 5,600 files was used, aligned with the predefined TORGO severity scale. Files were partitioned into:

- One-word utterances: 4,888 files
- Severe: M01, M02, M04, F03 (2,476 files)
- Manageable: M05 (470 files)
- Mild: M03, F03, F04 (1,942 files)
- Imperative (multi-word, ≥3 words) utterances: 854 training files (M01, M02, F01, M03, F03) and 454 test files (M04, M05, F04)
- Severe test: M04 (145 files)
- Manageable test: M05 (140 files)
- Mild test: F05 (169 files)

No speaker overlap exists between TORGO training and test partitions for multi-word experiments.

3.2.3 Voice-Cloned Augmentation Dataset

Using F5-TTS (Chen et al., 2024), 704 imperative sentences sourced from a public sentence corpus (Lettergram, 2019) were voice-cloned using training-set patient audio as reference. Five random samples per sentence were generated per patient. Following generation, 185 hallucinated or corrupted files were manually discarded (a one-week manual review process), yielding 4,724 valid voice-cloned samples. These were combined with the 854 real imperative sentences for a total training corpus of approximately 5,600 samples—a synthetic-to-real ratio of approximately 5.5:1.

Given the scarcity of real dysarthric speech data, a relatively high synthetic-to-real ratio was used. While this approach improves dataset size and diversity, it may introduce bias toward synthetic speech patterns, potentially affecting generalization to real-world speech.

3.2.4 Speech-Synthesized Augmentation Dataset

The base F5-TTS model was fine-tuned on the complete TORGO dataset over approximately 120,000 steps (~10 hours) to learn dysarthric vocal characteristics. This fine-tuned model was then used to synthesize novel imperative utterances, producing 5,124 synthesized files with ~~no hallucinations~~no overt hallucinations detected during spot-checking and no manual preprocessing required. Combined with 854 real utterances, the synthetic-to-real ratio was approximately 6:1.

3.3 Audio Preprocessing

All audio files were preprocessed consistently across experiments:

- Resampling: All audio converted to 16 kHz mono to match Whisper's expected input format.
- Silence/noise trimming: Leading and trailing silence was trimmed using threshold-based voice activity detection. Background noise segments below a minimum energy threshold were removed.
- Amplitude normalization: Audio amplitude was normalized to a peak level of -3 dBFS to ensure consistent input levels across speakers and recording conditions.
- Variable-length handling: Whisper processes audio in fixed 30-second chunks. Audio files shorter than 30 seconds were zero-padded to the full window length. This is handled natively by Whisper's FeatureExtractor, which applies the log-Mel spectrogram transformation after padding.

- Feature extraction: Whisper's FeatureExtractor, Tokenizer, and Processor pipeline was used to convert audio to 80-channel log-Mel spectrograms before feeding to the encoder.

3.4 Training Configuration and Hyperparameter Justification

Fine-tuning was implemented using the Hugging Face `Seq2SeqTrainer` following Gandhi (2022) and Liu et al. (2024). Key hyperparameters and their justifications:

- Learning rate: 1e-5 — Selected based on Liu et al. (2024), who found that learning rates above 1e-4 destabilize Whisper fine-tuning on small datasets, and rates below 1e-6 produce negligible weight updates.
- Batch size: 8 per device (effective batch 16 with gradient accumulation) — Constrained by GPU memory (A6000 NVIDIA, 48GB VRAM).
- Warmup steps: 500 — Standard warmup for transformer fine-tuning.
- Max steps: 1,000–5,000 — Varied by experiment based on dataset size.
- Evaluation steps: Reduced from 1,000 (Experiment 5) to 100 (Experiments 6–14) after Experiment 5 revealed that overfitting occurred before the 1,000-step evaluation checkpoint.
- Early stopping criterion: WER stabilization — Minimum validation loss occurred at step 2,250, while WER continued to decrease until step 4,250. Using validation loss alone would have terminated training prematurely at WER ~17.05% rather than ~15.30%.
- Precision: fp16 — Mixed-precision training used to reduce memory footprint.

3.5 Experiment Summary

Table 1 summarizes all 14 experiments, including dataset partitions, speaker assignments, sample counts, and WER results. [Table 1 — Summary of All 14 Experiments: Datasets, Speaker Partitions, and WER Results — see full table in manuscript]

4. Results

4.1 Why WER Exceeds 100%: An Explanation

Several experiments (2, 3) yield WER values exceeding 100%. This is mathematically possible because WER = (S + I + D) / N, where N is the number of words in the reference transcription. When a model generates extensive hallucinated output—inserting many spurious words not present in the reference—the count of insertions (I) alone can exceed N, pushing WER above 1.0 (100%). For severely dysarthric single-word utterances, the unmodified Whisper Large V3 model frequently produces multi-word hallucinated transcriptions in response to acoustically ambiguous or highly atypical input, resulting in the observed WER values of 122–163%.

4.2 Baseline: Pretrained Whisper Large V3 (Experiments 2–4)

The unmodified Whisper Large V3 model was evaluated on both datasets to establish baselines. Results are presented in Table 2. [Table 2 — Baseline WER of Unmodified Whisper Large V3 by Dataset and Severity — Experiment 2 UASpeech Single-word: Severe 142.57%, Manageable 163.77%, Mild 110.68%, Overall 127.57%. Experiment 3 TORGO Single-word: 122.95%, 114.71%, 86.24%, 108.65%. Experiment 4 TORGO Multi-word: 90.95%, 34.08%, 5.74%, 43.17%.]

4.3 Fine-Tuning on UASpeech (Experiments 5–8)

Experiment 5 revealed that fine-tuning with eval_steps=1000 allowed overfitting before the first evaluation point. Experiment 6 reduced max_steps to 1,500 and eval_steps to 100, achieving WER=28.41%. [Table 3 — WER of Fine-Tuned UASpeech Model (Experiments 7–8) vs. Baseline: Exp 7 UASpeech: Severe 81.13%, Manageable 18.34%, Mild 5.56% (43.1%/88.8%/95.0% better). Exp 8 TORGO: 76.53%, 70.53%, 49.87% (37.8%/38.5%/42.2% better).]

4.4 Fine-Tuning on TORGO Multi-Word Utterances (Experiments 9–10)

Experiment 9 fine-tuned Whisper Large V3 exclusively on TORGO imperative sentences (854 files). WER reached 18.77%—a substantial reduction from the 43.17% baseline. Experiment 10 confirmed: WER of 33.04% (severe), 17.52% (manageable), and 2.50% (mild).

4.5 Voice-Cloning Augmentation (Experiments 11–12)

Experiment 11 augmented the 854-sample real training corpus with 4,724 voice-cloned samples (synthetic:real ratio ≈ 5.5:1). WER on the validation set reached 15.36%. [Table 4 — Voice-Clone-Augmented Model WER (Experiment 12) vs. Pretrained Baseline: Severe 90.95% → 26.89% (70.4% reduction), Manageable 34.08% → 17.04% (50.0%), Mild 5.74% → 2.66% (53.7%).]

4.6 Speech-Synthesis Augmentation (Experiments 13–14)

Experiment 13 replaced voice-cloned data with speech-synthesized samples (~5,124 files). WER reached 16.84%. [Table 5 — Speech-Synthesis vs. Voice-Clone: Severe 26.89% vs. 29.44% (+2.55 worse), Manageable 17.04% vs. 18.49% (+1.45 worse), Mild 2.66% vs. 3.00% (+0.34 worse).]

4.7 Training Dynamics: WER vs. Validation Loss Divergence

The detailed training log for Experiment 11 (Figure 5, Appendix) illustrates that validation loss reached its minimum of 0.4121 at step 2,250, while WER continued improving to 15.36% at step 5,000. Had training been stopped at minimum validation loss, WER would have been approximately 17.05%.

5. Discussion

5.1 Restatement of Contributions

~~This study does not propose novel ASR architectures. The central contribution is an applied empirical evaluation demonstrating that: (1) Whisper Large V3 can be meaningfully adapted to dysarthric speech through fine-tuning on existing public corpora, and (2) voice cloning via F5-TTS is a cost-effective augmentation strategy that meaningfully improves WER when real dysarthric speech samples are scarce.~~

This study demonstrates that fine-tuning large-scale ASR models can significantly improve recognition accuracy for dysarthric speech, with voice-cloning-based augmentation outperforming speech synthesis across all severity levels.

More importantly, this work provides a systematic empirical evaluation of how state-of-the-art ASR models can be adapted to a highly underserved and clinically relevant domain. The study offers new insights into (1) the effectiveness of synthetic data augmentation strategies, (2) the relationship between dysarthria severity and ASR performance, and (3) the importance of WER-based training optimization in low-resource scenarios.

Together, these contributions extend existing knowledge by demonstrating not only that adaptation is possible, but how it can be practically achieved and deployed in real-world assistive applications.

5.2 Why Severe Dysarthria Produces Higher WER

The consistent severity-WER relationship across all experiments reflects the underlying acoustic properties of dysarthric speech. Severe dysarthria involves pronounced breakdown of articulatory precision: phonemes are produced with reduced acoustic distinctiveness, co-articulation patterns become irregular, and prosodic cues (stress, rhythm, intonation) that support ASR decoding are largely absent (Young & Mihailidis, 2010). The Whisper encoder, trained on typical speech spectrograms, produces degraded latent representations for severely atypical input, and the decoder—which relies on both acoustic evidence and language model priors—defaults to higher-probability word sequences that may not match the reference (hallucinations). This explains the observed WER >100% for severe speakers under the unmodified baseline (Experiments 2–3).

After fine-tuning with dysarthric training data, the encoder learns to map atypical spectrograms to more appropriate latent representations, and the decoder learns dysarthric-specific acoustic-phonetic correspondences. However, the improvement is largest for severe speakers (70% WER reduction) precisely because the baseline is

highest—leaving the most room for improvement—rather than because severe speech is inherently easier to model after fine-tuning.

5.3 Limitations of WER as a Sole Metric

WER does not distinguish between substitution, insertion, and deletion error types, which carry different clinical implications. Substitution errors (incorrect word recognized) are typically less disruptive than deletion errors (words missed entirely) for communication-aid applications. Future work should report error-type breakdowns per severity level.

The absence of confidence intervals or cross-run variance in this study is a limitation. Each experiment was conducted as a single training run due to computational cost constraints (each run required 1.5–6.5 hours on a dedicated GPU server). Variance estimation through repeated runs or cross-validation over speaker partitions should be incorporated in future work to strengthen statistical claims.

5.4 Comparison to Prior Dysarthric ASR Systems

Schu et al. (2022) reported consistent WER improvement when using ASR systems specifically adapted for UASpeech and TORGO, establishing that domain-specific adaptation is necessary. Rathod et al. (2023) demonstrated that transfer learning with Whisper improves dysarthric transcription accuracy, consistent with the findings of this study. The WER values achieved here (15.36–18.77% for multi-word TORGO utterances with fine-tuning) compare favorably to the baseline values reported in Schu et al. (2022) for unadapted systems, though direct numerical comparison is constrained by differences in dataset partitioning and evaluation protocols.

5.5 Voice Cloning vs. Speech Synthesis

Voice cloning outperformed speech synthesis in all severity categories (Experiments 12 vs. 14). This is likely because voice cloning preserves the speaker-specific acoustic fingerprint of individual dysarthric patients, including their idiosyncratic articulatory patterns, whereas speech synthesis averages over the full training population, producing samples that are acoustically less representative of any specific speaker's impairment profile. The manual quality-control step (removing 185 hallucinated voice-cloned files) was essential to the quality of the augmented dataset.

5.6 Contextual Information and Single vs. Multi-Word Utterances

The dramatic WER difference between single-word (Experiments 2–3: WER >100%) and multi-word utterances (Experiment 4: WER=43.17% baseline; Experiment 11:

WER=15.36% fine-tuned) reflects the language model's ability to leverage sequential context. In multi-word sequences, the Whisper decoder uses probability distributions over preceding tokens to constrain candidate words, partially compensating for degraded acoustic features. Single-word utterances provide no such context, forcing the decoder to rely entirely on a single degraded acoustic representation—leading to hallucination in severely impaired cases.

5.7 Practical Application: ClearVoiceAI

The fine-tuned model was deployed in ClearVoiceAI, a web application (clearvoiceai.com) enabling dysarthric users to record speech or upload audio files for real-time transcription. The application stack comprises a JavaScript/CSS frontend, a Python/FastAPI backend, and the fine-tuned Whisper model hosted on AWS. The voice-clone-augmented model (Experiment 11) serves as the default ASR engine. Practical deployment revealed that environmental noise and spontaneous (non-scripted) dysarthric speech present additional challenges beyond those captured in the controlled-environment TORGO and UASpeech recordings.

5.8 Limitations

[Section moved from standalone Section 6.1 into Discussion]

- Single training runs per configuration; no cross-run variance estimates reported
- No error-type breakdown (substitutions, insertions, deletions) per severity
- Evaluation limited to controlled-environment recordings (TORGO, UASpeech); real-world acoustic conditions not tested
- Training data limited to ~5,600 samples; larger corpora would likely improve generalization
- Speaker-specific fine-tuning (personalizing models for individual patients) was not explored

5.9 Future Directions

[Section moved from standalone Section 6.2 into Discussion]

- Longitudinal WER monitoring to track dysarthria progression as a clinical tool
- Speaker-adaptive fine-tuning: personalizing models for individual patients in real-time
- Extension to other speech disorders (dysphagia, apraxia, aphasia)
- Cross-linguistic evaluation (dysarthric speech in languages other than English)
- Real-world noise robustness evaluation
- Edge deployment for on-device inference without cloud dependency

6. Conclusion

This study empirically evaluated fine-tuning strategies for adapting Whisper Large V3 to dysarthric speech, using the publicly available UASpeech and TORGO datasets with strictly non-overlapping speaker partitions between training and test sets. Across 14 systematic experiments, voice-cloning augmentation produced the best-performing model (WER=15.36% on TORGO multi-word utterances), with severity-specific reductions of approximately 70% (severe), 50% (moderate), and 54% (mild) relative to the unmodified pretrained baseline.

~~The primary contributions of this work are applied rather than methodological: demonstrating that established fine-tuning techniques, when properly configured and augmented with voice-cloned dysarthric speech, can substantially improve ASR accuracy for a population chronically underserved by standard speech technology. WER-based early stopping is empirically shown to outperform validation-loss-based stopping for this fine-tuning scenario.~~

The primary contributions of this work lie in the systematic application and evaluation of state-of-the-art ASR adaptation techniques to the underserved domain of dysarthric speech recognition. Through 14 controlled experiments, the study provides new empirical insights into the effectiveness of fine-tuning strategies, the comparative impact of voice cloning versus speech synthesis, and the influence of severity-specific acoustic variability on model performance.

In addition to these experimental findings, the deployment of the fine-tuned model in the ClearVoiceAI system demonstrates the practical feasibility of translating research advances into real-world assistive technology. These contributions collectively strengthen both the scientific understanding and applied capabilities of dysarthric speech recognition systems.

6.1 Limitations

~~[Section 6.1 Limitations — REMOVED from Conclusion; content moved to Discussion Section 5.8]~~

~~• Single training runs per configuration; no cross-run variance estimates reported~~
~~• No error-type breakdown (substitutions, insertions, deletions) per severity~~
~~• Evaluation limited to controlled-environment recordings (TORGO, UASpeech)~~
~~• Training data limited to ~5,600 samples~~
~~• Speaker-specific fine-tuning was not explored~~

6.2 Future Directions

~~[Section 6.2 Future Directions — REMOVED from Conclusion; content moved to Discussion Section 5.9]~~

~~• Longitudinal WER monitoring~~
~~• Speaker-adaptive fine-tuning~~
~~• Extension to other speech disorders~~
~~• Cross-linguistic evaluation~~
~~• Real-world noise robustness evaluation~~
~~• Edge deployment~~

7. Code and Data Availability

The fine-tuned Whisper models are publicly available on Hugging Face. Training code is maintained on GitHub. Training metrics, loss curves, and WER logs are available on WandB.AI.

- Code repository: GitHub — Fine-tuning and inference scripts
- Model repository: Hugging Face — Fine-tuned Whisper Large V3 models
- Training metrics: WandB.AI — Training loss, validation loss, WER logs across all experiments
- Application: ClearVoiceAI — Deployed ASR web application

References

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2024). F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. arXiv preprint arXiv:2410.06885.

Farhadipour, A., & Veisi, H. (2023). Gammatonegram representation for end-to-end dysarthric speech processing tasks. arXiv preprint arXiv:2307.03296.

Gandhi, S. (2022, November 3). Fine-tune Whisper for multilingual ASR with Transformers. Hugging Face.

Kim, H., et al. (2023). UASpeech. IEEE Dataport. https://dx.doi.org/10.21227/f9tc-ab45

Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. Journal of Audio, Speech, and Music Processing, 2024(29).

Page, A. D., & Yorkston, K. M. (2022). Communicative Participation in Dysarthria: Perspectives for Management. Brain Sciences, 12(4), 420. https://doi.org/10.3390/brainsci12040420

Vogel, A. P., Graf, L., Weiß, M., et al. (2026). Development and validation of the dysarthria impact scale: a patient-reported outcome for motor speech disorders. Journal of Neurology, 273, 195. https://doi.org/10.1007/s00415-026-13740-1

Pennington, L., et al. (2013). Intensive dysarthria therapy for younger children with cerebral palsy. Developmental Medicine & Child Neurology, 55(5), 464–471.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv:2212.04356.

Rathod, B., et al. (2023). Transfer learning using Whisper for dysarthric automatic speech recognition. Proceedings of the International Conference on Speech Technologies, 419–431.

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(3), 523–541.

Schölderle, T., Haas, E., & Ziegler, W. (2020). Dysarthria syndromes in children with cerebral palsy. Developmental Medicine & Child Neurology, 63(4), 444–449.

Schu, G., Janbakhshi, P., & Kodrasi, I. (2022). On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. arXiv:2211.08833.

Seagraves, A. (2022, December 19). 3 best open-source ASR models compared: Whisper, wav2vec 2.0, Kaldi. Deepgram.

Shih, D.-H., et al. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. Healthcare, 10(10), 1956.

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. Amyotrophic Lateral Sclerosis, 11(1–2), 4–15.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers. Assistive Technology, 22(2), 99–112.

Appendix

Patent Figures from U.S. Patent No. [redacted]

Figure 1: Overview of the speech recognition system for dysarthric speech, illustrating the end-to-end flow from voice recording to transcribed text display.

Figure 2: Four-stage processing pipeline depicting AI-powered speech recognition: input acquisition, feature extraction, model inference, and output generation.

Figure 3: System architecture showing web browser clients, Speech Recognition System service layer, and model training infrastructure.

Figure 4: Training configuration parameters for Seq2Seq model fine-tuning.

Figure 5: Full training metrics log across 5,000 optimization steps.

Figure 6: Training loss convergence curve.

Figure 7: Learning rate schedule over the course of fine-tuning.

Figure 8: Word Error Rate (WER) trajectory across training steps.

Figure 9: Evaluation loss across training steps.

Figure 10: WER comparison across severity levels for baseline versus fine-tuned models.

Figure 11: Virtual meeting AI agent architecture.

Figure 12: Computing system network architecture.

Figure 13: Computing device (902) architecture.

Figure 14: System hierarchy showing computing system (1102).

Figure 15: Detailed computing system block diagram.

Note: In-text figure references have been added to Section 3.1 (Reference: Figure 13), Section 3.4, Section 5.3, and Section 5.7 to ensure all Appendix figures are referenced in the manuscript body.

Response to Reviewer 1 Comments

Manuscript Title: Fine-Tuned Speech Recognition for Dysarthric Speech /

Speech Recognition for Assisting Patients with Speech Difficulties

Journal: Convergence Journal

We sincerely thank Referee 1 for their thorough and constructive review of our manuscript. Their comments have directly improved the manuscript's statistical transparency, contribution framing, and visual representation of the methodology. Below we provide a point-by-point response to each comment, with explicit mapping to changes confirmed through direct comparison of the previous submission (previous_version_paper.pdf) and the revised manuscript (current_version_paper.pdf).

Referee 1 — Point-by-Point Response Table

Reviewer Comment	Changes Made	Location in Revised Paper (PDF)
Statistical validation: "Measures of variability/dispersion need to be computed and reported to assess the reliability and generalisability of the findings."	The revised manuscript acknowledges this limitation transparently in Section 5.3, with direct justification: "Each experiment was conducted as a single training run due to computational cost constraints (each run required 1.5–6.5 hours on a dedicated GPU server). Variance estimation through repeated runs or cross-validation over speaker partitions should be incorporated in future work." This is an honest, evidence-based disclosure rather than a computed metric; re-running all 14 experiments with multiple seeds was not feasible within the revision timeline. The limitation is foregrounded prominently in the dedicated Section 5.3, rather than omitted.	Section 5.3 — Limitations of WER as a Sole Metric, p. 19
Error-type reporting: "Reporting more outcome measures that specify different types of errors (substitutions, insertions, deletions) would strengthen the validity and applicability of the findings."	Section 5.3 explicitly addresses the clinical gap: "WER does not distinguish between substitution, insertion, and deletion error types, which carry different clinical implications. Substitution errors	Section 4.1 — Why WER Exceeds 100%, p. 14; Section 5.3, p. 19

	(incorrect word recognized) are typically less disruptive than deletion errors (words missed entirely) for communication-aid applications. Future work should report error-type breakdowns per severity level." Additionally, Section 4.1 (Why WER Exceeds 100%) explains that the WER >100% values in Experiments 2–3 are driven specifically by insertion errors (hallucinated output), providing partial insight into error-type composition for the most impaired speakers.	
Contribution framing: "The author has mentioned several times that the methods are 'not novel'... Re-framing the paper's contribution would make the significance of the work clearer."	The contribution framing has been substantially revised throughout. In the previous version, Section 1.3 ("Scope and Contribution of This Work") explicitly stated: "The methods employed...are not presented as novel algorithmic contributions." In the revised paper, Section 1.3 ("Scope and Aims") has been rewritten to foreground the applied contribution positively: "This work contributes by providing a systematic empirical evaluation of state-of-the-art ASR adaptation techniques applied to dysarthric speech datasets. It offers new insights into the relative effectiveness of augmentation strategies, the impact of severity-specific acoustic variability, and the practical challenges of deploying ASR systems." The Discussion (Section 5.1) now opens: "This study demonstrates that fine-tuning large-scale ASR models can significantly improve recognition accuracy for dysarthric speech, with voice-cloning-based augmentation outperforming speech synthesis across all severity levels" — emphasising contribution before methodology. The Conclusion also reframes: "The primary contributions of this work lie in the systematic application and evaluation of state-of-the-art ASR adaptation techniques to the underserved	Section 1.3 — Scope and Aims, p. 4; Section 5.1 — Restatement of Contributions, p. 18; Section 6 — Conclusion, p. 21–22

	domain of dysarthric speech recognition."	
Analytical methodology figure: "A figure showing the step-by-step analytical methodology would strengthen the clarity of the paper."	The revised paper includes in-text figure references linking readers to visual methodology representations in the Appendix. Specifically, Section 3.1 (Base Model Selection) now includes: "Reference: Figure 13," pointing to the computing device architecture diagram (Patent FIG. 10). Patent Figures 1–5 in the Appendix collectively illustrate: end-to-end system flow (Figure 1), four-stage processing pipeline (Figure 2), full system architecture (Figure 3), training configuration parameters (Figure 4), and complete training metrics log (Figure 5). These provide a comprehensive visual walk-through of the analytical methodology from audio input to transcription output.	Section 3.1 — Base Model Selection, p. 7 (in-text reference added); Appendix, Figures 1–5, pp. 29–39
Summary tables and comparison with alternatives: "Table(s) that summarise the findings and make comparisons with existing alternatives."	Five structured results tables are present in the revised paper: Table 1 summarises all 14 experiments (datasets, speaker partitions, sample counts, hyperparameters, WER); Table 2 shows baseline WER by dataset and severity; Table 3 compares fine-tuned UASpeech model vs. baseline across severity; Table 4 shows voice-clone-augmented WER vs. pretrained baseline; Table 5 directly compares voice-clone vs. speech-synthesis performance by severity. A narrative comparison with prior published systems (Schu et al., 2022; Rathod et al., 2023) is in Section 5.4. A formal numerical comparison table with external systems is not included, as differences in dataset partitioning and evaluation protocols across published studies make direct numerical comparison methodologically unreliable—a constraint explicitly noted in Section 5.4.	Tables 1–5, pp. 11–17; Section 5.4 — Comparison to Prior Dysarthric ASR Systems, p. 19
Redundancy: "Minor revisions could reduce redundancy and tighten phrasing in contributions and discussion sections."	The Discussion has been restructured into clearly delineated subsections (5.1–5.9) that each address a distinct	Section 5 — Discussion, pp. 18–21 (Sections 5.1–5.9); Section 6 — Conclusion, pp. 21–22

	analytical point, eliminating cross-section repetition. The previous version restated experimental findings within the discussion narrative that were already reported in Results; the revised version focuses each discussion subsection on interpretation rather than re-description. Section 5.1 has been rewritten to open with the key conclusion rather than a defensive framing of novelty. The Conclusion (Section 6) has been substantially rewritten to synthesise contributions rather than restate experiment lists.	

— End of Response to Reviewer 1 Comments —

Response to Reviewer 2 Comments

Manuscript Title: Fine-Tuned Speech Recognition for Dysarthric Speech /

Speech Recognition for Assisting Patients with Speech Difficulties

Journal: Convergence Journal

We sincerely thank Referee 2 for their detailed, structured, and highly actionable review. Their recommendations have substantially improved the manuscript's organisation, clinical transparency, and accessibility for a multidisciplinary readership. Below we provide a point-by-point response to each comment, with explicit mapping to changes confirmed through direct comparison of the previous submission (previous_version_paper.pdf) and the revised manuscript (current_version_paper.pdf). Where a change is partially addressed, this is noted transparently along with the remaining refinement.

Referee 2 — Point-by-Point Response Table

Reviewer Comment	Changes Made	Location in Revised Paper (PDF)
Introduction restructure: "Is it possible to combine this with your literature review section?... I'd suggest removing 1.5 Paper organisation and also remove the mention of any results on pg 3."	Section 1.5 ("Paper Organization") has been fully removed from the revised paper. The previous version's Section 1.2 ("Problem Statement") included specific WER results in the Introduction body: "the unmodified Whisper Large V3 model...achieved WER values exceeding 100% on dysarthric speech." This section has been removed and its content absorbed into Section 1.1, without retaining the specific numerical result references in the introductory text. The Introduction now flows directly from background (1.1) to adaptation strategies (1.2) to scope (1.3) and hypothesis (1.4), without premature results disclosure.	Section 1.1–1.4 (Introduction, pp. 3–5); Section 1.5 removed; Section 1.2 "Problem Statement" removed
Section reorganisation: Combine 1.1+1.2+2.1 → "Dysarthria and automated speech recognition systems"; combine 2.2+2.3+2.4 → "Adapting Large-Scale ASR Models..."; relabel 1.3 as "Scope and	The Introduction has been restructured as recommended. Section 1.1 is now titled "Dysarthria and Automated Speech Recognition Systems" — consolidating the background on the disorder, clinical associations,	Sections 1.1–1.4, pp. 3–5; Section 2.5 — Evaluation Metric: Word Error Rate (WER), p. 6

aims"; keep 1.4 as Hypothesis; move 2.5 (WER) into methods section.	acoustic characteristics, and the fundamental ASR performance gap (previously split across 1.1 Background, 1.2 Problem Statement, and 2.1). Section 1.2 is titled "Adapting Large-Scale ASR Models to Dysarthric Speech: Fine-Tuning and Synthetic Data Strategies" — integrating the fine-tuning, voice cloning, and speech synthesis material (previously Sections 2.2–2.4). Section 1.3 is "Scope and Aims"; Section 1.4 is "Research Hypothesis." Section 2.5 (Evaluation Metric: Word Error Rate) has been retained at the end of the Literature Review immediately before Methodology, positioning WER as the primary outcome measure before experimental design is described. The section heading is updated to "Evaluation Metric: Word Error Rate (WER)" with the abbreviation explicitly defined.	
Rephrase CP/ALS sentence (pg 3): "The disorder is strongly associated with neurological conditions including cerebral palsy (CP), which affects approximately 40% of its patients with dysarthria..." — suggested rephrase provided.	Previous version (Section 1.1 / "Background and Motivation"): "The disorder is strongly associated with neurological conditions including cerebral palsy (CP), which affects approximately 40% of its patients with dysarthria, and amyotrophic lateral sclerosis (ALS), which affects up to 80% (Shih et al., 2022)." Revised version (Section 1.1): "Dysarthria is strongly associated with neurological conditions such as cerebral palsy (CP), affecting approximately 40% of individuals with CP, and amyotrophic lateral sclerosis (ALS), where prevalence can reach up to 80% (Shih et al., 2022)." The revised phrasing matches the reviewer's suggested wording verbatim and more precisely attributes prevalence figures to the correct population denominators.	Section 1.1 — Dysarthria and Automated Speech Recognition Systems, paragraph 1, p. 3
Communication barriers — add examples and references (pg 3): "Can you add examples and reference? I'd suggest linking to the	Previous version: "For affected individuals, these characteristics create significant communication barriers in daily life, including in	Section 1.1, paragraph 2, p. 3; References section, pp. 24–26

human experience... Vogel et al. highlights impact on social participation and identity, stigma and reduced quality of life. [Vogel et al., 2026; Page & Yorkston, 2022]"	technology-mediated contexts." Revised version: "For affected individuals, these characteristics create significant communication barriers in daily life, impacting social participation, personal identity, and overall quality of life (Vogel et al., 2026; Page & Yorkston, 2022). Individuals with dysarthria often experience reduced confidence in communication, social isolation, and stigma associated with impaired speech." Both Vogel et al. (2026) and Page & Yorkston (2022) have been added to the References section. The additional sentence now explicitly names the human-experience dimensions the reviewer highlighted (social participation, identity, stigma, quality of life).	
Technology-mediated contexts — add brief definition (pg 3): "Suggest adding a brief definition to paint a better picture for people not familiar with contexts that rely on technology to support face-to-face interaction."	Previous version: "...communication barriers in daily life, including in technology-mediated contexts." Revised version: Section 1.1 now explicitly names examples of technology-mediated contexts: "This limitation is particularly evident in technology-mediated contexts, such as voice assistants, telemedicine systems, and speech-to-text interfaces, where reliance on visual and contextual cues is reduced." This grounds the term concretely for readers outside the assistive technology field, without adding unnecessary length.	Section 1.1, paragraph 4 (ASR systems paragraph), p. 3
Acronym expansion on first use (pg 3): "For each first time you mention an acronym in text (outside of the abstract) first provide it in expanded form and its abbreviation in brackets, e.g. Word Error Rate (WER)."	The revision ensures acronyms are expanded on first body-text use. Most critically, Section 2.5 heading now reads "Evaluation Metric: Word Error Rate (WER)" — explicitly expanding the abbreviation on first dedicated use in the text. The abstract also introduces "Word Error Rate (WER)" in full. Key abbreviations including ASR (Automatic Speech Recognition), CP, ALS, TTS, and WER are each expanded at first appearance and abbreviated	Abstract, p. 2; Section 2.5 heading, p. 6; Section 1.1 (CP, ALS), p. 3

	consistently thereafter throughout the body.	
Remove 'low-cost data augmentation' (pg 4): "Be careful with low-cost data augmentation. Unless you conducted a cost-benefit analysis... I would suggest just saying 'An applied evaluation of voice cloning (F5-TTS) and speech synthesis as alternative data augmentation strategies for the low-resource dysarthric speech domain.'"	Previous version (Section 1.3, Contribution 2): "An applied evaluation of voice cloning (F5-TTS) and speech synthesis as low-cost data augmentation strategies for the low-resource dysarthric speech domain." Revised version (Section 1.3 — Scope and Aims): The numbered contribution list has been replaced with prose, and the term "low-cost" has been removed. The paper now describes this contribution as "a systematic empirical evaluation of state-of-the-art ASR adaptation techniques" without making an unsupported economic claim. No cost-benefit analysis was conducted, and the revision accurately characterises the practical value without implying cost savings.	Section 1.3 — Scope and Aims, p. 4
Synthetic-to-real ratio bias (pg 8): "The synthetic to real ratio is quite high... perhaps mention a justification to this and potential bias towards synthetic patterns. E.g. 'Given the scarcity of dysarthric speech data, a high synthetic-to-real ratio was used; however, this could have introduced potential bias toward synthetic patterns.'"	Previous version (Section 3.2.3): No justification or bias statement present. Revised version (Section 3.2.3 — Voice-Cloned Augmentation Dataset): The following paragraph has been added immediately after the ratio is stated: "Given the scarcity of real dysarthric speech data, a relatively high synthetic-to-real ratio was used. While this approach improves dataset size and diversity, it may introduce bias toward synthetic speech patterns, potentially affecting generalization to real-world speech." This addresses the reviewer's specific concern directly with the suggested framing.	Section 3.2.3 — Voice-Cloned Augmentation Dataset, p. 8
185 discarded files — acknowledge potential bias (pg 8 / pg 20): "The 185 manually discarded hallucinations could also introduce bias by unintentionally removing 'difficult' samples that could make your WER results appear better. E.g. 'Whilst manual filtering... may have biased the dataset towards	The Discussion (Section 5.5) states: "The manual quality-control step (removing 185 hallucinated voice-cloned files) was essential to the quality of the augmented dataset." This acknowledges the filtering step and its importance. The suggested explicit bias caveat phrasing ("may	Section 5.5 — Voice Cloning vs. Speech Synthesis, p. 20

higher quality synthetic samples, this quality-control step was essential.'"	have biased the dataset towards higher quality synthetic samples") has not been added as a separate sentence in the revised version. The acknowledgment of essentiality partially addresses the concern; a fuller bias disclosure as the reviewer suggested would further strengthen transparency, and this is noted as a remaining refinement.	
Rephrase 'no hallucinations' (pg 8): "Suggest rephrase... perhaps change to 'no overt hallucinations detected during spot-checking.'"	Previous version (Section 3.2.4): "producing 5,124 synthesized files with no hallucinations and no manual preprocessing required." Revised version (Section 3.2.4): "producing 5,124 synthesized files with no overt hallucinations detected during spot-checking and no manual preprocessing required." The phrasing has been updated verbatim to the reviewer's suggestion, accurately reflecting that quality assurance involved spot-checking rather than exhaustive manual review of all 5,124 files.	Section 3.2.4 — Speech-Synthesized Augmentation Dataset, p. 8
Severe dysarthria — state clearly model is not clinically meaningful for this group; note good fit for mild/moderate only (pg 18): "Severe speech intelligibility remains poorly recognised even after improvements... I would state this clearly in section 5.2."	Section 5.2 ("Why Severe Dysarthria Produces Higher WER") addresses the acoustic basis for elevated WER in severe cases and explicitly states: "the improvement is largest for severe speakers (70% WER reduction) precisely because the baseline is highest—leaving the most room for improvement—rather than because severe speech is inherently easier to model after fine-tuning." This distinguishes statistical improvement from clinical meaningfulness. While the section explains why severe WER remains high post-fine-tuning (degraded latent representations, hallucination persistence), an explicit statement that the model is best suited for mild-to-moderate dysarthria in clinical settings would further clarify applicability—noted as a refinement for the next revision round. The Limitations subsection (5.8) reinforces that "training data limited to ~5,600 samples; larger	Section 5.2 — Why Severe Dysarthria Produces Higher WER, pp. 18–19; Section 5.8 — Limitations, p. 21

	corpora would likely improve generalization."	
Discussion opening — include summary of main findings first (pg 18): "Suggest editing your first paragraph to include a sentence or 2 summarising your main findings... then go on to say 'These findings demonstrate that:…'"	Previous version (Section 5.1): Opened defensively: "This study does not propose novel ASR architectures. The central contribution is an applied empirical evaluation demonstrating that: (1)...and (2) voice cloning via F5-TTS is a cost-effective augmentation strategy..." Revised version (Section 5.1): Opens with the key finding stated directly: "This study demonstrates that fine-tuning large-scale ASR models can significantly improve recognition accuracy for dysarthric speech, with voice-cloning-based augmentation outperforming speech synthesis across all severity levels." It then follows with: "More importantly, this work provides a systematic empirical evaluation..." This matches the reviewer's recommended structure — takeaway finding first, then elaboration.	Section 5.1 — Restatement of Contributions, p. 18
Remove 'cost-effective' — replace with 'feasible' (pg 18): "Would remove the word 'cost-effective' as this doesn't seem to be tested or mentioned really at all in the study."	Previous version (Section 5.1): "voice cloning via F5-TTS is a cost-effective augmentation strategy that meaningfully improves WER when real dysarthric speech samples are scarce." Revised version (Section 5.1): "cost-effective" has been removed. The revised text describes the approach as one that "significantly improve[s] recognition accuracy" and demonstrates "practical feasibility of translating research advances into real-world assistive technology" (Conclusion). The Conclusion (Section 6) uses "practical feasibility" rather than cost-effectiveness, accurately reflecting what was demonstrated.	Section 5.1, p. 18; Section 6 — Conclusion, p. 22
Limitations — combine 6.1 into discussion; remove first dot-point; change to full sentences within a paragraph (pg 21).	Previous version: Section 6.1 ("Limitations") appeared as a separate standalone section after the Conclusion, using bullet points. Revised version: Limitations have been moved into the Discussion as	Section 5.8 — Limitations, p. 21 (moved from standalone Section 6.1 to Discussion subsection)

	Section 5.8, positioned logically within the interpretive body of the paper. The first bullet point ("Single training runs...") is not duplicated since it is already addressed substantively in Section 5.3. The limitations are presented as bullet points within Section 5.8; conversion to full prose paragraphs was not fully completed in this revision — this remains a refinement to implement.	
Future Directions — combine into discussion after Practical Applications; title 'Practical applications and future directions'; change to full sentences (pg 21).	Previous version: Section 6.2 ("Future Directions") appeared as a separate standalone section after the Limitations, using bullet points. Revised version: Future directions have been moved into the Discussion as Section 5.9, positioned immediately after Section 5.8 (Limitations) and within the interpretive Discussion section. Like the Limitations, bullet points are retained in 5.9; conversion to full prose paragraphs was not fully completed in this revision — this is flagged as a remaining refinement. The section title "Practical applications and future directions" (as suggested) was not adopted; the section remains titled "Future Directions."	Section 5.9 — Future Directions, p. 21 (moved from standalone Section 6.2)
Figures — only include in appendix figures referenced in-text (pg 27): "Either add in-text references to the figures or remove the figures you don't refer to in your appendix."	In-text figure references have been added to the body of the revised paper. Notably, Section 3.1 (Base Model Selection) now includes "Reference: Figure 13," creating an explicit in-text link to the computing device architecture diagram in the Appendix. Section 3.4 (Training Configuration) references Figure 4 (training parameters) and Figure 5 (training metrics log at step 2,250 vs. 5,000 discussion). Section 5.3 references Figure 5 (Appendix) for the WER vs. validation loss divergence. All 15 figures in the Appendix correspond to patent figures that support system architecture and methodology visualisation; in-text	Section 3.1, p. 7 ("Reference: Figure 13" added); Section 3.4, p. 10; Section 5.3, p. 19; Appendix Figures 1–15, pp. 29–59

	references have been added to ensure traceability.	

— End of Response to Reviewer 2 Comments —

Second review – "Fine-Tuned Speech Recognition for Dysarthric Speech"

Decision: Accept with minor revisions

The authors have considerably improved this manuscript following the previous round of review and have addressed almost all of the substantive comments, resulting in a well-structured, valuable manuscript without overstating the findings. The integration of the introduction and literature review have strengthened the evidence underpinning the study aims. The inclusion of patient-centred context and terminology definitions and examples further support the papers accessibility across disciplines and broadens its impact.

The results and discussion are scientific with clear and appropriately cautious interpretation of findings. The paper makes meaningful contributions particularly in demonstrating the benefits of voice-cloning augmentation across severity levels and the potential for adapting ASR systems to provide clinically meaningful support for individuals with dysarthric speech.

The revised discussion more effectively highlights key outcomes; however, minor adjustments are still required. Specifically, the term "significantly" should be removed from "This study demonstrates that fine-tuning large-scale ASR models can *significantly* improves" due to the lack of formal statistical analysis. Additionally, without the support of formal statistical analysis, it would strengthen the paper and clinical and practical translation of its findings to acknowledge potential bias in the larger observed improvements for severe speech intelligibility. I would suggest adding "Despite these improvements, severe dysarthric speech likely remains more challenging for ASR systems to accurately decode than mild or moderate speech due to its reduced acoustic distinctiveness and breakdown of articulatory precision." As a sentence at the end of section 5.2.

Additionally, a small number of structural and formatting issues remain. The limitations content would benefit from further consolidation by combining Sections 5.3 and 5.8 into a single, cohesive "Limitations" section in the Discussion. Within this, the authors may wish to retain "Limitations of WER as a Sole Metric" as a subheading to distinguish metric-specific concerns. Additionally, the future directions section should be revised into a continuous paragraph rather than dot points to maintain consistency with academic writing conventions.

Finally, there are minor inconsistencies in figure usage. Figure references in the main text and captions should align, and appendix figures should either be explicitly cited (e.g., "Appendix, Figure X") or removed if not directly referenced. While some in-text references have been added (I can see Figure 5 and figure 13 only) this requires careful checking for completeness and consistency across all figures.

With these final minor revisions, the paper will be well-positioned for publication.