Fine-Tuning whisper model for ASR for Dysarthria

Omri Benbenisty
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
omri.benbenisty@mail.huji.ac.il

Tomer Cohen
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
tomer.cohen13@mail.huji.ac.il

Abstract

Automatic Speech Recognition (ASR) systems often struggle with the accurate transcription of speech from individuals with dysarthria, a motor speech disorder that impairs articulation. Here, we explore the potential of fine-tuning pre-trained Whisper models to improve ASR performance on dysarthric speech. Utilizing the TORGO dataset, we fine-tuned Whisper models of various sizes using a well-established fine-tuning framework. We evaluated the performance of these models using two key metrics: Word Error Rate (WER) and Character Error Rate (CER). Our results reveal a significant reduction in error rates, with the fine-tuned Whisper-small model achieving a CER of 18% on the test set, a substantial improvement over the 67% CER observed with the corresponding pre-trained model. These findings suggest that fine-tuning Whisper models can substantially enhance the accuracy of ASR systems for dysarthric speech, providing a pathway toward more inclusive speech recognition technology. Our trained model and code are freely available at: https://github.com/tomerco4/FinetuneDysarthria

1 Introduction

Speech recognition technology has seen rapid advancements in recent years, achieving remarkable accuracy in transcribing typical spoken language (Radford et al., 2023; Kriman et al., 2020; Gulati et al., 2020; Baevski et al., 2020). However, these advancements have predominantly focused on typical speech, leaving behind individuals with speech disorders, such as dysarthria. Dysarthria is a motor speech disorder resulting from neurological damage, affecting the muscles involved in speech production. Consequently, individuals with dysarthria experience varying degrees of speech impairment, making their speech difficult to understand for both humans and ASR systems. Developing specialized ASR systems capable of accurately transcribing dysarthric speech is therefore imperative (Liu et al., 2021), as these individuals often lack alternative communication methods such as keyboards or touchscreens (Hosom et al., 2003). There have been several attempts to create such systems, using methods such as fine-tuning pre-trained models and data-augmentation, (Qian et al., 2023; Shahamiri et al., 2023; Javanmardi et al., 2024; Almadhor et al., 2023; Zheng et al., 2023; Yu et al., 2023), but the task of ASR on dysarthric speech still remains a challenge.

The Whisper model, developed by OpenAI, is a state-of-the-art ASR system (Radford et al., 2023). Released in September 2022, Whisper is a pre-trained model that directly benefits from large-scale labeled audio-transcription data, unlike its predecessors, such as Wav2Vec 2.0 (Baevski et al., 2020), which were pre-trained on unlabeled audio data. Whisper's extensive pre-training on labeled data enables it to excel in speech-to-text tasks with minimal additional fine-tuning. Architecturally, Whisper is a Transformer-based encoder-decoder model, functioning as a sequence-to-sequence system that maps audio spectrogram features to text tokens. The process begins with converting raw audio inputs into a log-Mel spectrogram, which the Transformer encoder processes to generate a sequence of hidden states. The decoder then autoregressively predicts text tokens, conditioned on both the previous tokens and the encoder hidden states.

Whisper has demonstrated robust performance on a wide range of speech tasks. However, like most ASR systems, its performance diminishes when confronted with non-standard speech patterns, such as those present in dysarthria. To address this limitation, we investigated the effectiveness of fine-tuning Whisper models to better accommodate dysarthric speech. Using the TORGO dataset (Rudzicz et al., 2012), which contains speech samples from individuals with dysarthria, we fine-tuned multiple Whisper models of varying sizes and assessed their performance using Word Error Rate (WER) and Character Error Rate (CER) as evaluation metrics.

This paper presents the methodology and results of our fine-tuning process, highlighting the impact of model size on performance. Our findings indicate that fine-tuning significantly improves ASR accuracy for dysarthric speech, particularly with the Whisper-small model. These results underscore the potential of tailored ASR systems in enhancing communication for individuals with speech disorders.

2 Dataset

We decided to use the TORGO dataset, a comprehensive database designed for the study of dysarthric speech. Initially, we considered the "Dysarthria and Non-Dysarthria Speech Dataset" from Kaggle but found it unsuitable due to numerous missing data samples and inconsistencies, such as mismatches between sound samples and text prompts, and inclusion of instructions rather than transcriptions in the text prompts.

The TORGO database includes both audio recordings and 3D articulatory features from speakers diagnosed with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), two of the most common causes of dysarthria (Kent & Rosen, 2004). The dataset contains recordings from both dysarthric and non-dysarthric individuals. However, for our task we focused exclusively on the dysarthric speech data.

The dataset comprises 3,323 paired audio and text samples from 3 female and 5 male speakers with dysarthria.

2.1 Pre-processing

Each sample in the dataset is recorded using either a head-mounted microphone or an array microphone, or both. To minimize background noise, we prioritized head-mounted microphone recordings when available. We excluded any samples flagged as unsuitable by the dataset authors (e.g., text prompts containing 'xxx'), as well as samples where the text prompt was an image or included instructions for the participant. Additionally, we removed any non-transcription clarifications from the text prompts. This pre-processing resulted in a total of 3,150 samples for a total of 3.12 hours.

2.2 Data Splitting

Given the relatively small size of our dataset, we wanted to reduce the risk of overfitting by carefully splitting the data into training, validation, and test sets. To ensure that the model's performance on the test set would not be influenced by overfitting to specific individuals, we designated all samples from a single subject, "M01", as the test set. We selected this specific subject because he contributed 371 samples, accounting for approximately 12% of the dataset (0.40 hours).

For the remaining data, we randomly shuffled the samples and allocated approximately 12% to the validation set, resulting in 2,389 samples for training (2.32 hours) and 390 samples for validation (0.39 hours).

3 Metrics and Baselines

3.1 Metrics

To evaluate the performance of the models, we used two metrics: Word Error Rate (WER) and Character Error Rate (CER). We included CER in addition to WER due to the presence of numerous single-word samples in the dataset, which could lead to misleading WER values. CER provides a more nuanced assessment by being more forgiving in cases where the predicted and correct text are nearly identical or when the pronunciation of two different words is similar.

3.2 Baselines

For baseline comparisons, we evaluated the performance of all pre-trained Whisper models available on HuggingFace, which includes Whisper-tiny, Whisper-base, Whisper-small, Whisper-medium, and Whisper-large-v3 (Radford et al., 2023). Each model has a different number of parameters, and we assessed their WER and CER on the test set (Table 1).

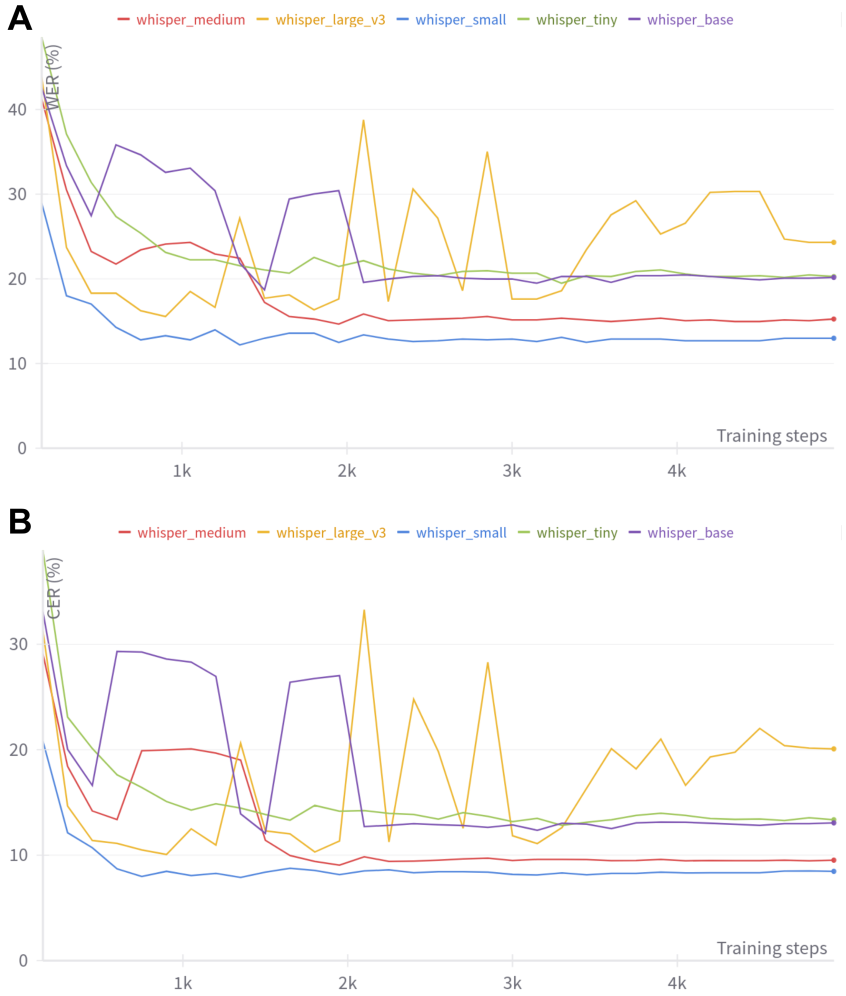

Figure 1: Validation performance during training. **A.** WER (%) for all of the five fine-tuned models. **B.** CER (%) for all of the five fine-tuned models.

4 Fine-tuning

We fine-tuned each of the five pre-trained Whisper models and compared their performance to that of the original models. Our fine-tuning process was largely based on the "Fine-Tune Whisper For Multilingual ASR with Transformers" tutorial from HuggingFace, which was adapted from a transcription task for Hindi to suit our specific problem.

The first step in the fine-tuning process was ensuring that all audio samples were resampled to a frequency of 16kHz, as required by the Whisper model. We then extracted log-Mel spectrograms from the audio files, which serve as the input for the Whisper model.

The models were fine-tuned using the cross-entropy objective function. We monitored the models' performance during training using the Weights & Biases package (Biewald, 2020), logging WER and CER on the validation set. Several hyperparameters were optimized, including batch size, learning rate, weight decay, and optimizer choice. For the final models, we selected the AdamW optimizer with a weight decay of 0. For the Whisper-tiny, Whisper-base, and Whisper-small models, we used a batch size of 16 and a learning rate of 1e-5. Due to memory constraints, we reduced the batch size for Whisper-medium and Whisper-large-v3 to 8 and 4, respectively, and lowered the learning rate to 1e-6.

After evaluating the validation CER during training (Fig.1), we chose to use the models after 3,000 training steps for inference and evaluation of the test set.

5 Results

Fine-tuning resulted in significant improvements in both CER and WER on the test set (Table 1 and Figure 2). Notably, all fine-tuned Whisper models outperformed their pre-trained counterparts, with the Whisper-small model achieving the highest overall performance. Specifically, the Whisper-small model achieved a WER of 26.89% and a CER of 17.97%, compared to the pre-trained model's WER of 77.97% and CER of 67.40%.

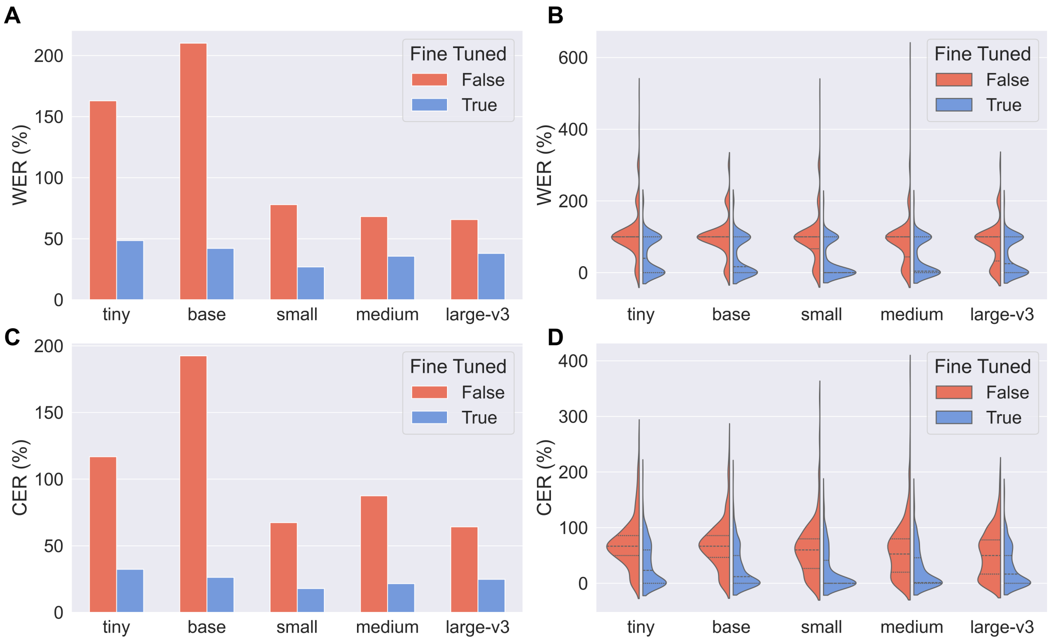

Figure 2: Test set results for all of the five pre-trained and fine-tuned models. **A.** Average WER (%). **B.** WER (%) distribution (We removed cases with WER > 700 for better visualization). **C.** Average CER (%). **D.** CER (%) distribution (We removed cases with CER > 700 for better visualization).

	WER		CER		
Model	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Parameters (M)
Tiny	162.96	48.49	116.80	32.34	39
Base	210.15	42.12	192.52	26.30	74
Small	77.97	**26.89**	67.40	**17.97**	244
Medium	68.25	35.74	87.52	21.58	769
Large	65.77	38.01	64.27	24.84	1550

Table 1: Average WER and CER (%) on the test set.

Interestingly, the best-performing model was not the one with the highest number of parameters, which could be attributed to the small size of our training dataset. This finding suggests that simpler models may be more effective in certain scenarios, particularly when dealing with limited data. Overall, fine-tuning Whisper on the TORGO dataset yielded more accurate ASR models, demonstrating performance comparable to state-of-the-art ASR models on standard speech tasks (e.g., Whisper's WER of 29.2% on the Multilingual ASR task (Radford et al., 2023)).

In several cases, the fine-tuned model achieved perfect transcriptions (60% for the small model, Fig. 3), even for challenging audio recordings, highlighting the potential to use our model for dysarthtic ASR in real world applications. We also conducted a qualitative analysis of specific audio examples from the test set (Table 2).

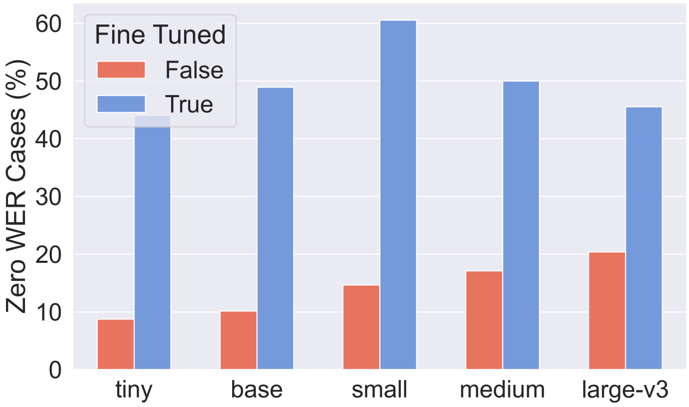

Figure 3: Perfect transcription results for the test set.

Original Text Prompt	Fine-tuned	Predicted Text	WER (%)	CER (%)
she wore warm fleecy woolen overalls		you are a worm sweetie what in the world	150.00	69.44
	V	you wore warm fleecy woolen overalls	16.67	8.33
before thursdays exam review every formula		before their days they can have a review and read from you	166.67	69.05
	V	before thursdays exam review every formula	0	0
this is not a program of socialized medicine		it is not a program of your canada afirge sorry	62.50	52.27
	V	the program of slaughter likes medicine	75.00	50.00

Table 2: Test set examples for the fine-tuned and pre-trained whisper-small models.

6 Discussion

Our study demonstrates the significant potential of fine-tuning Whisper models to improve ASR performance on dysarthric speech. The results underscore the value of model adaptation for non-standard speech patterns.

One of the surprising results of our experiments is the Whisper-small model's strong performance relative to larger models, such as Whisper-medium and Whisper-large-v3. This result might suggests that, when dealing with smaller, more specialized datasets like TORGO, simpler models can outperform more complex counterparts. This may be due to the larger models' tendency to overfit the limited training data. Although, these results might be an outcome of the smaller batch size and training strategies we used for the bigger models.

Another key takeaway from our study is the impact of dataset quality and size on model performance. The TORGO dataset, while carefully curated, remains relatively small by ASR standards. Further improvements may require larger and more diverse datasets of dysarthric speech, perhaps incorporating additional languages, dialects, or speech disorders. Synthetic data augmentation techniques, such as generating dysarthric speech variants from typical speech using voice conversion models, could also help mitigate the challenge of limited training data.

Despite the progress demonstrated in our study, practical challenges remain. ASR systems are increasingly being integrated into everyday devices and services, including smartphones, smart home devices, and virtual assistants. Ensuring that these systems perform reliably for people with dysarthria in real-world scenarios—such as in noisy environments, with spontaneous speech, or under different microphone conditions—requires further testing and validation.

7 Future Work

There are several directions for future research based on our findings. First, investigating additional input types for the model, such as integrating acoustic features, phonetic transcriptions, or articulatory data.

Second, given the limited availability of dysarthric speech data, it will be interesting to explore data augmentation techniques using regular speech data. For instance, generating dysarthric-like speech from people with regular speech using techniques such as voice conversion or synthetic speech manipulation could help expand the dataset and improve model robustness. This would also allow for a more diverse and comprehensive training set without requiring significantly more data collection from individuals with speech impairments.

Furthermore, we leave for future work the task of training larger Whisper models (e.g., Whisper-medium and Whisper-large-v3) with larger batch sizes. In our current study, the computational resources available limited the batch size for these models, potentially contributing to their underperformance. We anticipate that increasing batch sizes, along with further hyperparameter tuning, could unlock the full potential of these larger models.

Finally, it will be interesting to investigate the application of our fine tuned Whisper model in real-world settings, evaluating its performance in noisy environments and across different types of dysarthria.

References

Ahmad Almadhor, Rizwana Irfan, Jiechao Gao, Nasir Saleem, Hafiz Tayyab Rauf, and Seifedine Kadry. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications*, 222:119797, 2023.

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

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

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

J-P Hosom, Alexander B Kain, Taniya Mishra, Jan PH Van Santen, Melanie Fried-Oken, and Janice Staehely. Intelligibility of modifications to dysarthric speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pp. I–I. IEEE, 2003.

Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech. *IEEE Journal of Biomedical and Health Informatics*, 2024.

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

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

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

Zhaopeng Qian, Kejing Xiao, and Chongchong Yu. A survey of technologies for automatic dysarthric speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):48, 2023.

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

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

Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

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

Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai. Improving the efficiency of dysarthria voice conversion system based on data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:4613–4623, 2023.