Enhancing Whisper: Transfer Learning for Dysarthric Speech Recognition

Mirella Günther

Thesis committee: Zhengjun Yue, YuanYuan Zhang, Thomas Durieux

Key Terms

- **Dysarthria:** A neuromotor speech disorder characterised by weakened, slowed, or poorly coordinated muscle movements involved in speech production
- **Automatic Speech Recognition (ASR):** Technology that converts a speech signal into the corresponding sequence of words
- **Transfer Learning:** Using knowledge gained from solving one problem and applying it to a related but different problem

Problem Description

- Dysarthria usually goes hand-in-hand with additional physical impairments
- This means that individuals with dysarthria cannot interact with technology through traditional means, e.g. screens, keyboards
- Therefore, voice technology can increase their quality of life
- However, ASR achieves lower accuracies for people with dysarthria

Research Question

How can automatic speech recognition models designed for typical speech be adapted using fine-tuning to better recognise dysarthric speech?

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

The Whisper Model

- Whisper [1] is a state-of-the-art ASR model created by OpenAI
- There exist seven variants of the model of various sizes, of which the **Large-V2** and **Large-V3** models will be used
- Large-V2 has 1550M parameters and was trained on 680,000 hours of speech data
- Large-V3 also has 1550M parameters but was trained on 1 million hours of weakly labelled audio and 4 million hours of audio labelled by Whisper Large-V2 [2]
- The **distil-Whisper** [3] models are versions of the Whisper models that have been reduced to be more time- and space-efficient
- The distilled versions of Large-V2 and Large-V3 have 756M parameters each

Transfer Learning Methodology

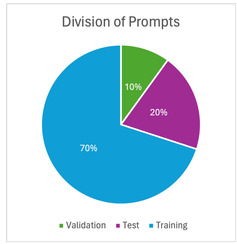
- Transfer learning involves taking the pre-trained Whisper and distil-Whisper models, which have demonstrated proficiency in typical speech recognition, and further training them with dysarthric speech data
- **Low-Rank Adaptation** (LoRA) will be used to perform parameter-efficient fine-tuning
- **Word Error Rate** (WER) is the chosen evaluation metric. This is calculated as follows:

$$\text{WER} = \frac{\text{Substitions} + \text{Deletions} + \text{Insertions}}{\text{Number of Words}} \times 100\% \quad (1)$$

The TORGO Dataset

- Eight participants with dysarthria
- Age- and gender-matched control participants
- Collaboration between the University of Toronto and the Holland-Bloorview Kids Rehab Hospital in Toronto [4]

To create this dataset, the participants were given prompts to read out loud. The majority of these prompts were the same between participants, so there is a **high degree of repetition** within the data. Therefore, the test, training and validation sets were created based on the prompts.



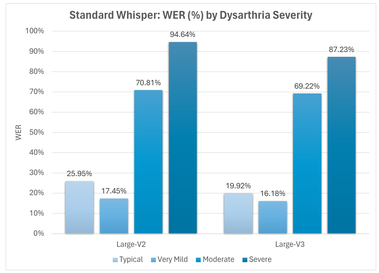
Experiments

💬 **Zero-shot Testing:** Run the test set against the Whisper and distil-Whisper models to establish performance before fine-tuning

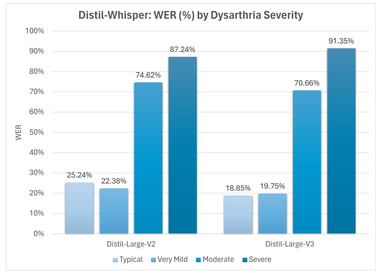
⚙️ **Fine-tune & Evaluate Whisper:** Use the training sets to train the chosen Whisper models using Low-Rank Adaptation (LoRA)

🔊 **Distil-Whisper Models:** Fine-tune and evaluate the distil-Whisper models to establish whether they can perform comparably

💬 Zero-Shot Testing Results



(a) Baseline WER (%) by Dysarthria Severity for Standard Whisper Models



(b) Baseline WER (%) by Dysarthria Severity for Distil-Whisper Models

- The two best-performing models in the zero-shot testing were the Large-V2 and Large-V3 models, with WERs of 67.36% and 63.28% on the entire dysarthric test set
- The typical speech is recognised worse than 'very mild' dysarthric speech, potentially because dysarthric subjects articulate more carefully
- In a breakdown by severity, the WER usually increases as the severity of dysarthria increases
- The distilled models perform comparably to the standard Whisper models, and the distil-Large-V2 model even outperforms the Large-V2 model for 'severe' dysarthric speech

⚙️ Fine-Tuning Results

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

Figure 2. Relative Difference in WER Before and After Fine-Tuning

- All models have achieved improved WER scores after being fine-tuned on dysarthric data
- Aside from the Large-V3 model, all models have improved their performance on dysarthric data but achieved higher WERs on typical and/or 'very mild' dysarthric speech
- The standard Large-V2 model is outperformed by the distil-Large-V2 model, because of the overfitting to 'severe' dysarthric speech and the production of more hallucinations
- The Large-V3 model outperforms the distil-Large-V3 model, suggesting that the larger amount of parameters allows it to better capture the dysarthric speech features
- Even though the distilled Whisper models performed worse on severe dysarthric speech than their non-distilled counterparts, the fine-tuning took 21.93% (Large-V2) and 17.24% (Large-V3) less time

Main Contributions

- LoRA fine-tuning generally improved the recognition of dysarthric speech
- The best-performing model was the Large-V3 model, with a relative WER reduction of 22.65%
- The distil-Large-V3 model performed worse than the standard Large-V3 model, but the distil-Large-V2 model outperformed the Large-V2 model
- The fine-tuning on the distilled models was quicker, so distilled Whisper models can provide a good alternative for scenarios with limited computational resources

References

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

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

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

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

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

mirellagunther@student.tudelft.nl