Task 6

Propose a model self-supervised learning pipeline to cater dysarthric speech and describe how you would do continuous learning in 500 words.

Introduction

Dysarthric speech is characterised by irregular articulatory patterns resulting from neurological injury of the motor component of the motor-speech system. This pattern of speech poses a major challenge to Automatic Speech Recognition (ASR) models that were trained on typical speech data. Through self-supervised learning (SSL), followed by continuous learning, we can leverage large amounts of unlabelled dysarthric speech audio to ensure our model learns the different kinds of speech and ensure model robustness.

Data Preprocessing and Augmentation

The first step in building an SSL pipeline is to preprocess our dataset. This dataset will mainly include unlabelled dysarthric speech data from sources such as audios taken from YouTube or crowdsourced speech recordings taken from healthcare settings or rehabilitation platforms.

Audio samples should be converted into a consistent format (16kHz PCM, mono-channel format) using tools such as FFMPEG, to standardise the data. To improve on speech clarity in the audio, noise-reduction models such as DeepFilterNet can be used to eliminate irrelevant background noise. For audio recordings that are long, it should be segmented into manageable 20-second chunks to optimise model processing. Additionally, filters can be used to isolate speech-only segments, removing irrelevant non-speech portions such as music or silences. Lastly, data augmentation techniques, such as random cropping, can be applied to ensure diversity in our training data.

SSL Pre-training

Self-supervised pre-training allows our model to learn meaningful speech representations without needing labelled data. We can use Wav2Vec2 to learn directly from the audio waveforms. Time masking technique can also be used to mask random segments of the input audio during training to encourage our model to predict the masked portions, ensuring model's robustness. The model is then trained using contrastive loss functions such as InfoNCE, getting the model to distinguish between similar and dissimilar speech features.

Fine-Tuning for Dysarthric Speech

After getting our pre-trained model, we will fine-tune it on a small, labelled dataset of dysarthric speech (that is open source) such as TORGO or UASpeech. We will first freeze the lower layers of the model and train only the output layers to adjust to the target task. Next, we will unfreeze all layers and fine-tune the entire model. To assess the performance of the fine-tuned ASR model, Word Error Rate (WER) can be used as the primary evaluation metric. WER measures the accuracy of the model's transcriptions by comparing the predicted text to the corresponding ground-truth text.

Continuous Learning Strategy

Continuous learning ensures that our ASR model remains relevant and adapts to new speech patterns over time. As such, new data will have to be incorporated so that our model can learn continuously. We can collaborate with healthcare institutions to collect new dysarthric speech recordings from real-world applications such as patients with this disorder. These audios can be used to fine-tune the model periodically. Apart from new data, old speech samples can be reused to ensure that our model does not forget previously learned information.

Conclusion

This SSL pipeline utilises large volumes of unlabelled data and robust pre-training techniques to develop a generalizable ASR model. Continuous learning further enables the model to adapt to evolving speech patterns over time.