Self-supervised Learning for Automatic Dysarthric Speech Recognition

Dysarthria is defined as a speech disorder resulting from neurological injury and is characterized by poor articulation of phonemes. Dysarthric speech represents a challenge for ASR systems due to slurred speech, inconsistent articulation of phonemes and high inter-speaker variability. In this essay, a pipeline for training an ASR system for dysarthric speech is proposed, as well as a strategy for continuous learning.

Dataset

Two of the most widely used dysarthric speech dataset is the UASpeech dataset and the TORGO dataset, which consists of labelled transcripts and speech recorded from both dysarthria and non-dysarthria speakers. Due to the absence of other larger scale datasets for dysarthric speech, we may consider collecting more recordings (with permission) from dysarthria patients in natural settings such as during clinical sessions or within home environments. These recordings do not need to be transcribed, and can be used for self-supervising pre-training.

Data Preprocessing

Similar to Karima et al., we convert all audio to 16KHZ PCM and use a Voice Activity Detection (VAD) model to filter and remove the long silences. Given the distinct nature of silence vs non-silence audio, we should be able to adapt the VAD model to detect the start of dysarthric speech. Furthermore, due to the irregular nature of dysarthric speech, we can also utilize an Audio Event Detection (AED) model to distinguish speech vs non-speech events such as background noise. The AED can then be used to remove non-speech segments.

Self-Supervised Pre-training

Following the approach of Karima et al., we first perform self-supervised pre-training on unlabelled dysarthric audio using Lfb2vec. This method masks random segments of the input log-Mel features, then learn to predict the masked positions of the resulting context vectors using a contrastive loss function, optimizing it to be similar to the unmasked target vector in the same position. We can also use a flatNCE contrastive loss function, which provides stability with smaller batch sizes and non-homogeneous data, such as those in dysarthric audio samples.

Supervised Finetuning

After self-supervised pre-training, the model is finetuned on labelled dysarthric audio datasets, such as the UASpeech and the TORGO dataset. Similar to

Karima et al, we can do this in two stages. In the first stage, we only finetune the linear projection layer and in the second stage, we train the entire model.

Continuous Learning

To improve the accuracy of our model over time, we can use continuous learning to retrain our model using new data. A subset of the new data can be randomly selected to be transcribed. The un-transcribed audio can be added to the self-supervised pre-training dataset, while the transcribed audio can be added to the supervised finetuning dataset. In order to prevent catastrophic forgetting, we use a mixture of old and new data to retrain the model using a small learning rate. After retraining, we validate the new model on a held-out dataset. This process allows us to adapt to new patients with different levels of dysarthria.