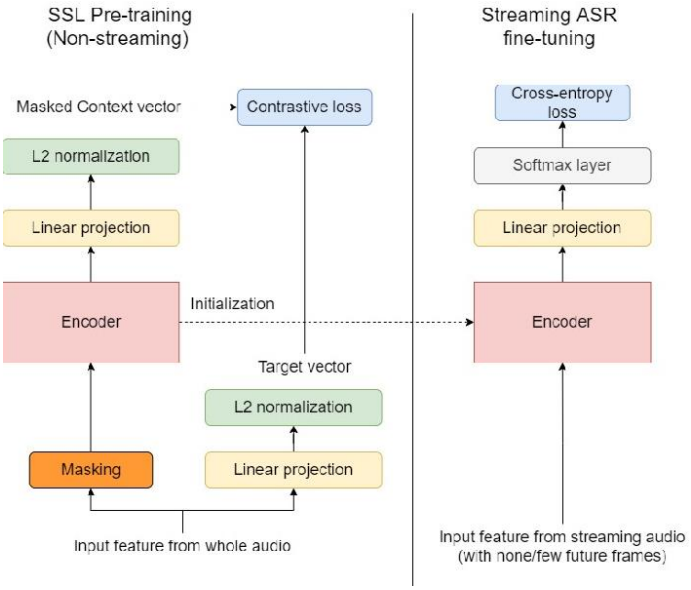
HTX xData Q6

Dysarthric speech is described as slurred or slow speech that can be difficult to understand. Due to muscle or nerve damage, the speech may be slurred or mumbling, or short and choppy bursts with several pauses. There are several datasets with dysarthric speech, such as UASpeech database, which consists of dysarthric speakers and normal controls reading isolated words.



The SSL pre-training pipeline is as shown above. To make use of the pipeline described in the journal article, we could pass the log-Mel features of dysarthric speech into the encoder instead of masking the audio, while passing the corresponding log-Mel features of speech from normal controls through the linear projection layer. Since the dysarthric speech and normal speech are reading the same words, the contrastive loss forces the encoder to pick up appropriate features from the dysarthric speech to generate the appropriate context vector. Given the lack of dysarthric speech in the wild, this will also help augment our dataset, as we may use positive and negative examples to improve the encoder's accuracy. Next, for the finetuning process, we train the ASR model to understand dysarthric speech by training on the text labels of the audio.

The journal article suggests the use of a Voice Activity Detection (VAD) Model and Audio Event Detection (AED) model before the audio is passed to the encoder to improve the ASR accuracy. The VAD model may not need much finetuning, while differentiating between intentional and unintentional speech in the AED model could improve accuracy as well. However, this may require manual labelling.

For continuous learning, we may deploy the model and allow dysarthric speakers to provide feedback on the text predictions generated by the model. We can then pass the text prediction through a linear projection to generate a context vector, while passing the recorded dysarthric speech through the encoder to generate another context vector. We can then use contrastive loss to further train the model, where positive feedback provides positive examples, while negative feedback provides negative examples to the SSL pipeline. The text-to-vector linear projection may be trained using contrastive loss against normal speech.

In terms of the lack of datasets for dysarthria, we may pre-train the encoder on data from dysarthric speakers in other languages. The encoder may be able to learn features from dysarthric speech that is common regardless of language. To augment the dataset, we may also try mixing in normal speech that has been distorted with features of dysarthria, such as masking parts of the speech, changing the speech volume or pace, or adding unintentional noise to the audio, to pass through the encoder in the SSL pipeline.