Generie ASR models might not work well on dysarthric speech as it can have different characteristics compared to non-dysarthric speeches and there can be high variance among different dysarthric speeches. Furthermore, labeled dysarthrie speech data may be limited.

Adapting from the model proposed in the paper, SSL pipeline for dysarthrie speech include 3 main components: data pre-processing, pre-training and fine-tuning. They are mostly similar to the model proposed in the paper but with small adaptation to dysarthric speech in the pre-training phase.

Within the data pre-processing pipeline, data will be prepared in a format that enhances model performance. Similar to the paper, this includes converting data to 16kHz 16bit PCM, removing long silences, segmenting audio with maximum length of 20 seconds, extracting log-Mel features and passing them through the AED model to separate speech from background noise. Speech-filter, speech-cropand Rand-crop will be applied. A potential addition to the pre-processing method proposed by the model to adapt to dysarthrie speech includes normalizing the tempo and pitch of the speeches before removal of long silences.

During pre-training, Lfb2vec is used to optimize flatNCE between masked context vectors and target vectors. The prior is obtained after masked log-Mel features pass through encoder, linear projection layer and L2 normalization layer while the latter is obtained after log-Mel features pass through linear projection and L2 normalization layer. This is also similar to the model proposed in the paper. However, since limited dysarthric speech data are available, using non-dysarthric data can improve model performance since they share similar characteristics (e.g. same language but with variations in speed, volume, tempo, speed or pitch). Hence, we can pre-train with both dysarthric and non-dysarthric data where they share the same encoder but with a catered linear projection layer. Under the assumption of limited availability of labeled data for dysarthric speech, single-head SSL might work better. Should more data be available, we can explore multi-head SSL or supervised learning.

Finally, the fine-tuning portion consists of a 6-layer LC-BLSTM and linear projection layer similar to the paper. After which the results are then passed into a 5-gram Language Model as decoder to obtain the transcribed results.
To perform continuous learning, it is important that more data can be collected so that the model is able to learn from the new data to improve its performance. Varying the types of dysarthric speech allows for better model generalization. To make use of these data for continuous learning, fine-tuning should be performed regularly using both the existing and newly collected data.