Two lessons from this paper can be applied to Continual Learning (CL) for dysarthric speech. First, it is shown that separate per-language output heads are beneficial for multi-lingual SSL. For dysarthric CL, this motivates initialising a new output head when fine-tuning on the dysarthric speech. Second, it is shown that a pre-trained offline model can be fine-tuned for mismatched streaming mode, while still matching the performance of streaming pre-training from scratch. For dysarthric CL, this gives hope for further dysarthric fine-tuning from a non-dysarthric seed model.

There are several design choices in the paper that may not be optimal for dysarthric speech. First, the utterance segmentation uses a one second threshold applied to the detected silence duration. However, dysarthric speech may express longer and more varied intra-sentential silence durations and frequencies. As such, this simple silence threshold may not be appropriate, and data-driven methods may instead be warranted. Second, the use of contrastive learning does not agree with more recent empirical observations that discrete classification criteria tend to yield better downstream model performance than continuous regression criteria, such as contrastive learning, even with the use of both positive and negative examples. As such, regardless of how the seed model was pre-trained, the output layer and criterion should be replaced with a discrete classification criterion, such as cross-entropy with masked prediction, when at the dysarthric CL fine-tuning stage.

Here is a proposed pipeline for CL on dysarthric speech, with the aim of downstream application to speech recognition.

Data preparation
Dysarthric speech data is difficult to obtain, because of the more limited population of speakers. Reliable annotated dysarthric data is even more expensive to obtain, due to the difficulty of transcribing dysarthric speech. As such, a cost balance can be optimised by sourcing for as much dysarthric speech-only data as possible, and only relying on limited open-source annotated dysarthric data or labelling a limited portion of the speech data.
The speech needs to be segmented, to fit within available GPU memory. Instead of using the threshold detection approach from the paper, it is instead proposed to use a small neural network cut-point detector that is trained on a limited portion of manually segmented dysarthric speech.
A limited portion of this segmented speech is then manually transcribed. Transcription should be done after segmentation, so that segment alignment between words and speech is known. Dysfluencies should be annotated, to ensure close matching between the annotation and speech for easier modelling with limited data.

SSL CL
Start from an open-source pre-trained seed model. Use HuBERT stage2 style SSL, by computing K-means cluster targets of dysarthric speech from the hidden embeddings of the seed model. Prevent catastrophic forgetting by 1) using LoRA, 2) train output layer alone first before training LoRA weights, and 3) regularise by constraining parameter distance from seed model.

Speech recognition fine-tuning
Fine-tune the model for dysarthric speech recognition on limited annotated data. Use a hybrid framework. The lexicon should comprise pronunciation dysfluencies. Use minimum Bayes' Risk training and decoding for better generalisation.