The paper explained a pipeline about how to use uncurated data to build a streaming ASR model, and shows that the performance of the model with SSL and in domain data.

The pipeline could be described as below:

1. Collect unlabelled dysarthric audio data through recording.

2. Label part of the data for benchmarking ASR performance.

3. Data Pre-processing for Uncurated Dysarthric Audio: Collecting high-quality dysarthric data is a challenge, hence we need to utilize the uncurated audio data following the same way in the paper. Use VAD and AED to filter the audio data and hence ensures the SSL model focuses on relevant acoustic information.

4. Pretraining: Building upon the demonstrated in-domain performance of Lfb2vec SSL pre-training combined with flatNCE contrastive loss using uncurated in-domain data, we propose extending this approach to further enhance its effectiveness. The core idea is to leverage the robust self-supervised representations learned by Lfb2vec with flatNCE on readily available, uncurated in-domain data, and then explore how these pre-trained representations can be efficiently adapted or fine-tuned for a broader range of downstream tasks within the same domain, even with limited labeled data. This extension aims to establish a more versatile and resource-efficient framework for achieving strong in-domain performance across various applications, minimizing the reliance on extensive manual data curation or large annotated datasets.

5. Finetuning: After pre-training, the learned representation could serve as a powerful feature extractor for ASR tasks. At this stage, label a small amount of dysarthric speech, we could tune the model to fit for dysarthric speech.

Continuous Learning:

After the initial deployment of our Automatic Speech Recognition (ASR) model, we can implement a robust continuous learning pipeline that leverages real user interactions to consistently enhance performance. This strategy focuses on incrementally updating the model, particularly the Lfb2vec encoder, which is crucial for learning effective speech representations.

We can continuously feed massive amounts of unlabeled data generated by real user speech into the same self-supervised learning (SSL) pipeline used during the initial training phase. This allows the Lfb2vec encoder to constantly refine its understanding of speech patterns and acoustic nuances specific to our users' context.

In parallel, we will maintain a systematic process for continuously labeling new data. This newly labeled data will serve two critical purposes: fine-tuning the ASR model to improve its accuracy on specific linguistic patterns and benchmarking its performance against established metrics. To prevent catastrophic forgetting—where the model loses knowledge of previously learned patterns—we'll employ a fine-tuning with replay strategy. This involves

periodically re-training the model on a carefully selected subset of older, diverse labeled data alongside the new data.

By continuously learning from uncurated unlabeled data to update the speech encoder and systematically fine-tuning and benchmarking with new labeled data through replay, we can ensure the ASR model consistently improves its results and stays optimized for real-world usage. This iterative process allows for frequent model redeployments, ensuring our users always benefit from the latest advancements in speech recognition accuracy.