Dysarthria is a motor speech disorder that presents challenges for Automatic Speech Recognition (ASR) systems due to its heterogeneous acoustic patterns. Traditional supervised ASR models struggle with dysarthric speech due to limited availability of labelled datasets. The authors of the reference paper demonstrate that SSL can be effective in utilising vast amounts of untranscribed, noisy audio data on pre-train hybrid ASR models.

SSL Pipeline Design

1. Data Collection and Preprocessing:

- Data Sourcing:
    - Collect dysarthric speech from clinical sessions, support groups, public datasets and synthetic sources (e.g., voice conversion tools to simulate dysarthria).
- Preprocessing:
    - Standardise all audio into a uniform format (16 kHz, 16-bit PCM) and extract 80-dimensional log-Mel spectrograms.
    - Given the variability in dysarthric speech, apply Voice Activity Detection (VAD) to segment audio into speech/non-speech regions.
    - Introduce a Dysarthria-Specific Filtering
        - Adapt the paper's Audio Event Detection (AED) model to identify dysarthric speech characteristics (e.g., irregular pitch, slow articulation) and filter non-representative samples and ensure that only relevant segments are retained.
- Augmentation:
    - Introduce variability via time-stretching, dynamic time warping, and additive noise to mimic real-world dysarthric speech variations.

2. Self-Supervised Pre-training:

- Adopt an SSL approach inspired by Lfb2vec.
- Mask segments of the log-Mel features using a random, overlapping mask strategy tailored to shorter, irregular speech patterns typical in dysarthria.
- Feed the masked features into a robust encoder network (e.g., a six-layer BLSTM or a Transformer with multi-head attention) to learn context-rich representations.
- Optimize using the flatNCE contrastive loss to distinguish between masked and unmasked portions, using negative samples drawn from the same utterance to capture local temporal dependencies.

3. Supervised Fine-tuning:

- If available, fine-tune the pre-trained model on a smaller, labeled dysarthric speech dataset.
- Use a two-stage fine-tuning process:
    - Firstly, update only the final classification layers while freezing the encoder, and then fine-tune the entire network. This helps bridge any mismatch between pre-training on generic dysarthric features and the nuances of dysarthric speech recognition.

Continuous Learning Strategy:

- Deploy the model in a real-world system where new dysarthric speech data is continuously streamed.
- Implement a periodic re-training schedule that integrates new unlabeled data using SSL alongside a replay buffer of previously seen examples.
- Use online learning techniques such as elastic weight consolidation to mitigate catastrophic forgetting.
- Adjust the learning rate dynamically using schedulers similar to AdamW's warm-up and decay strategies.

- Monitor performance on a dedicated validation set and incorporate feedback loops that allow users to provide corrections, so that the model can be refined iteratively.