Task 6 Essay

To develop a self-supervised learning (SSL) pipeline for dysarthric speech, we can adapt the framework presented in the paper while addressing the unique challenges of dysarthric speech, which contains irregular articulation and variable speech patterns. The pipeline will consist of data pre-processing, SSL pre-training, and continuous learning.

The first step is data pre-processing. First, we can collect a large dataset of dysarthric speech from diverse sources, such as clinical recordings, assistive communication devices, and publicly available datasets like UA-Speech or TORGO. The audio will be converted to 16 kHz, 16-bit PCM format, and segmented using Voice Activity Detection (VAD) to remove long silences, with a max segment length of 20 seconds. An Audio Event Detection (AED) model that is trained to distinguish dysarthric speech from background noise and non-speech events can filter out irrelevant segments. Additional filters, such as "Speech-crop" and "Rand-crop" can help to isolate speech portions and augment the data. Given the variability in dysarthric speech, the AED model might need finetuning to better classify atypical speech patterns.

During the self-supervised learning pre-training stage, we can employ the Lfb2vec framework just like the paper, which masks log-Mel features and uses a bidirectional LSTM encoder to learn representations. For dysarthric speech, the contrastive loss function (eg. flatNCE) may be particularly beneficial due to its robustness to small batch sizes and unbiased mutual information estimation. The model will be pre-trained on unlabeled dysarthric speech data, with negative samples drawn from the same utterance to capture dysarthria-specific features. Multi-head multilingual SSL could also be explored if dysarthric speech data from multiple languages is available.

To adapt the model to new dysarthric speech patterns over time, we can implement continuous learning by periodically updating the model with new unlabeled data. This involves:

- Data Streaming: New audio data will be processed through the same pre-processing pipeline, with the AED model updated to handle new speech patterns.
- Incremental Training: The SSL model can be finetuned on new data using a reduced learning rate to avoid catastrophic forgetting. Techniques like elastic weight consolidation could preserve knowledge from earlier data.
- Feedback Loop: Transcribed dysarthric speech if available could be used to validate and refine the model. This ensures that it generalises to new speakers and severity levels.
- Evaluation: The model will be finetuned on a smaller labeled dysarthric dataset using a hybrid ASR architecture, similar to the paper's approach. Performance can be measured on test sets with varying dysarthria severity, focusing on Word Error Rate (WER) improvements over baseline models.

By combining data pre-processing, SSL pre-training, and continuous learning, this pipeline will enable adaptive ASR for dysarthric speech even with limited labeled data. The key lies in

leveraging uncurated data and contrastive learning to capture the unique characteristics of dysarthria while continuously refining the model with new inputs.