Essay question – Propose a model self-supervised learning pipeline to cater dysarthric speech and describe how you would do continuous learning

The paper presents a comprehensive approach to applying self-supervised learning (SSL) to automatic speech recognition (ASR), particularly emphasizing its application on uncurated, real-world audio datasets. The authors examine various aspects of SSL, including data preprocessing, choice of optimizer and learning rate scheduling, comparison of contrastive loss functions, and multiple pre-training strategies to improve streaming hybrid ASR models.

For developing a self-supervised learning pipeline to cater to dysarthric speech, we can draw insights from the paper and propose the following steps:

Data Preprocessing Pipeline: Given the nature of dysarthric speech, which may contain irregularities and significant variation, an advanced audio event detection (AED) model is crucial. The pipeline should include voice activity detection to filter out long silences, segmentation into manageable chunks, and feature extraction focused on capturing the nuances of dysarthric speech patterns.

Custom SSL Pre-Training (Dys2vec): Inspired by the Lfb2vec approach for SSL pre-training, we would develop 'Dys2vec,' a custom pre-training strategy tailored to dysarthric speech. Given the particular characteristics of dysarthric speech, the model should be trained to recognize patterns unique to such speech impairments. Using negative sampling, we would fine-tune the contrastive loss function to be sensitive to the irregularities present in dysarthric speech.

Contrastive Loss Functions: Adapting and possibly combining the loss functions mentioned in the paper (InfoNCE and flatNCE) could be beneficial. A loss function that accounts for the variability in speech among dysarthric patients will be crucial. This may include a loss function that penalizes deviations from common dysarthric patterns while still allowing for individual patient variability.

Cross-Lingual and Cross-Modal Learning: Given that dysarthric speech may share features across languages and modalities, a cross-lingual and cross-modal SSL could be beneficial. This would involve pre-training on a diverse set of languages and modalities, then fine-tuning on dysarthric speech data to leverage shared representations.

Continuous Learning: Dysarthric speech can change over time due to disease progression, therapy, or other factors. A continuous learning approach, potentially through online learning or regular model updates with new data, would ensure that the ASR system remains accurate and

sensitive to these changes. Techniques such as Elastic Weight Consolidation could be used to prevent catastrophic forgetting during continuous learning.

Evaluation and Iteration: Rigorous testing with dysarthric speech datasets will be crucial to evaluate model performance. Iterative refinements based on performance metrics, user feedback, and new data incorporation will be necessary to ensure the model adapts to the complexities of dysarthric speech.

Deployment and Real-World Testing: Deploy the model in a controlled environment initially to gather real-world data and user feedback, which will be invaluable for further model refinements.

The proposed pipeline should focus on creating a robust model that can handle the variability of dysarthric speech while being flexible enough for continuous adaptation and improvement. The goal is to improve accessibility and communication for individuals with speech impairments, and this approach provides a promising pathway to achieve that.