Dysarthric Speech Recognition: Self-supervised learning pipeline with Continuous learning

Dysarthria is a motor speech impairment characterized by poor rhythm, intelligibility and articulation. This results in people with dysarthria having difficulties communicating with Speech Recognition Technologies. This essay proposes a new pipeline for self-supervised learning in Automatic Speech Recognition (ASR) that is designed to address the challenges in dysarthric speech recognition with continuous learning functionality.

1. **Data Pre-processing Framework**
   The pipeline begins with robust pre-processing for dysarthric speech. Firstly, we will convert all audio 16 kHz 16-bit PCM format and implement specialized Voice Activity Detection (VAD) calibrated for irregular speech patterns. Followed by segmenting audio based on pauses, with flexible maximum lengths to accommodate for slower speeches. FInally, we extract 80-dimensional log-Mel features with extended processing windows and deploy a modified Audio Event Detection (AED) model trained on dysarthric speech characteristics to differentiate speech from background noise.

2. **Self-supervised Learning Architecture**
   By building upon the Lfb2vec architecture in the paper, the proposed model will use the 6-layer Bidirectional LSTMs with increased hidden dimensions of 800, and apply a targeted masking strategy with lower probability of about 0.05 to preserve more context. It will also implement flatNCE contrastive loss function to overcome the InfoNCE limitations and lastly, to establish robust base representations, we would initially train on a combination of non-dysarthric and dysarthric speech.

3. **Multilingual Multi-head approach**
   This approach is necessary to address speech diversity across the dysarthria types. FIrstly, implement a multi-head approach where the encoder is shared across different dysarthria severity levels. Next, create a severity-dependent projection layers: mild, moderate and severe. Lastly, allow the model to learn shared representations while accounting for severity-specific characteristics

This would be my proposed continuous learning framework that would operate through three synchronized mechanisms:

1. **Data collection and Quality control**
   We would have to establish an ongoing data collection pipeline from users with consent and implement automated quality assessment using the AED mode. Next, we would have to create data pools categorized by dysarthria type and severity. Split the data for

training and evaluation and this would help in maintaining a golden set of manually verified examples for evaluation

2. **Incremental Model Updates**
   For the model updates, I propose using an AdamW optimizer with scheduled learning rates, maye starting at 5e-4. These updates should be scheduled regularly with adaptive frequency based on data accumulation. Next, we should perform lightweight fine-tuning on new data while preserving previously learned representations and also implement privacy techniques to ensure user data protection.

3. **Evaluation and Model Selection**
   Create a comprehensive evaluation framework with metrics for intelligibility, naturalness, and WER. Benchmark against domain-specific datasets representing various dysarthria types. Implement A/B testing for model candidates before deployment and maintain versioned models to support rollback if needed

The proposed pipeline would significantly improve ASR for dysarthric speech by leveraging pre-training on both in-domain dysarthric speech and out-domain non-dysarthric speech. The continuous learning approach ensures adaptability to individual speech patterns while maintaining generalization across users, ultimately providing more inclusive speech technology for people with dysarthria.