Dysarthric Speech Self-Supervised Learning Pipeline

To effectively handle dysarthric speech without compromising performance on standard speech, a hybrid and continuous learning approach is proposed using three models.

Model 1 utilizes the Wav2Vec2 Large Pretrained Model for general speech recognition. Fine-tuning this model on dysarthric speech could degrade its performance on non-dysarthric speech; therefore, it should focus exclusively on standard speech while providing predictions accompanied by confidence scores. During its forward pass, the model produces logits representing unnormalized probabilities over the vocabulary at each time step. These logits help calculate the confidence level of each word or phoneme in the predicted sequence. The confidence score is defined as the maximum probability among all possible words or phonemes. If any word's confidence score falls below a certain threshold (e.g., 0.5), the entire sentence is passed to the generative language model (Model 3) for correction. Alternatively, the flagged words or sentence parts can be extracted and corrected using Model 3.

Model 2 is a specialized dysarthric speech model focusing exclusively on dysarthric speech patterns. It is fine-tuned on a dataset specifically curated from dysarthric speakers, enabling it to learn the nuances of dysarthric speech. This specialization allows the model to handle challenging pronunciations and variations not well-represented in typical speech datasets.

Model 3 incorporates a generative language model, such as GPT or BERT, to enhance recognition when dysarthric speech or low-confidence predictions occur. It analyses sentence context to predict missing words or correct low-confidence words using its contextual understanding. This model is particularly useful for out-of-vocabulary words or complex dysarthric pronunciations.

The confidence score mechanism and hybrid decision-making process involve Model 1 producing an initial prediction with associated confidence scores. For any words where the confidence is below the threshold, the whole sentence is passed to Model 3 for context-based correction. If Model 3 successfully identifies the correct word based on context, the entire sentence is saved as a new training sample for Model 2. This creates a feedback loop that enhances Model 2's understanding over time.

A continuous learning approach is adopted to improve Model 2 incrementally. As users, especially those with dysarthria, interact with the system, sentences with low-confidence predictions are flagged. These sentences, corrected either by human intervention or Model 3, are stored in a training database and periodically used to fine-tune Model 2. Incremental fine-tuning on newly collected dysarthric data allows Model 2 to gradually improve in handling new speech patterns, minimizing the risk of overfitting or losing performance on previously learned patterns. Employing a small learning rate ensures that this fine-tuning does not disrupt the model's existing features.

Data augmentation techniques like pitch shifting, noise addition, and time-stretching are used to enhance the diversity of dysarthric speech data. This helps Model 2 generalize better to a broader range of dysarthric speech patterns and environmental conditions. An adaptive curriculum learning strategy is also applied by starting fine-tuning with simpler, more clearly spoken dysarthric examples and gradually increasing complexity as the model improves. This ensures the model remains stable and improves incrementally without losing accuracy.

Regular validation and re-evaluation are essential. Model 1 and Model 2 are periodically evaluated on separate test sets for dysarthric and non-dysarthric speech. Model 1 should continue performing well on standard speech, while Model 2 should progressively improve on dysarthric speech without introducing errors. Techniques like elastic weight consolidation (EWC) are employed to ensure Model 2 retains knowledge from previous fine-tuning sessions.

By implementing this hybrid and multi-model approach, the system can effectively handle dysarthric speech while maintaining strong performance on general speech recognition tasks. Continuous learning ensures that Model 2 becomes increasingly specialized for dysarthric speakers without losing its general capability.