Task 6: SSL Essay

Proposed Self-Supervised Learning Pipeline for Dysarthric Speech and Continuous Learning

1. **Data Collection and Pre-processing:**

    I. **Data Acquisition:** Gather a diverse set of dysarthric speech recordings, ensuring representation across different severity levels and speech patterns. The dataset should cover various languages and dialects to account for the potential diversity in dysarthric speech.

    II. **Audio Event Detection (AED):** Implement an **Xception-based Audio Event Detection (AED)** model to filter out non-speech segments from audio data, such as background noise or irrelevant sounds. This is to enhance the quality of the training dataset by ensuring that only relevant speech segments are used for further processing.

2. **Feature Extraction and Representation Learning:**

    I. **SSL Model Selection:** Use a **self-supervised model** like wav2vec 2.0 or HuBERT, pre-trained on large-scale speech datasets.

    II. **Fine-tuning on Dysarthric Speech:** Fine-tune the pre-trained SSL model on the curated dysarthric speech dataset to adapt the model to the specific acoustic characteristics of dysarthric speech. This ensures the model learns representations that are more relevant to dysarthric speech type.

3. **Loss Function and Training:**

    I. **Contrastive Loss Function:** Use a **flatNCE** (flat noise-contrastive estimation) loss function, which addresses the drawbacks typically found in the standard InfoNCE loss.

    II. **Optimizer:** Use the **AdamW optimizer** with which has been shown to provide more stable convergence and better performance compared to traditional optimizers.

    III. **Learning Rate Scheduling:** Using 10% of steps as warm-up to reach the max learning rate and the rest of 90% with linear decay to reach set learning rate.

4. **Hybrid ASR Integration:**

I. **Model Architecture:** Integrate the fine-tuned SSL model into a **hybrid ASR (Automatic Speech Recognition)** framework. This hybrid model combines the benefits of SSL with conventional ASR methods, improving recognition accuracy for dysarthric speech.
II. **Training:** Train the hybrid ASR model using the representations learned from the SSL model. Optimize for metrics like Word Error Rate (WER) and Character Error Rate (CER) to measure improvements in the model's ability to transcribe dysarthric speech accurately.

5. **Evaluation and Iterative Improvement:**

I. **Performance Metrics:** Evaluate the model using standard ASR metrics, such as WER and CER, focusing on dysarthric speech performance. Track improvements over baseline ASR systems.
II. **Pre-training Strategies:** Incorporate **in-domain, multi-head multilingual SSL pre-training**, which has been shown to yield better results for hybrid ASR systems. This will allow the model to learn a more generalizable set of representations, which will improve its performance on dysarthric speech across different languages or dialects.

6. **Continuous Learning Strategy:**

I. **Incremental Updates:** Continuously incorporate new dysarthric speech data into the training pipeline. As more data becomes available, the model can be updated periodically to adapt to new speech patterns and variations in dysarthric speech.
II. **Avoiding Catastrophic Forgetting:** Use techniques like rehearsal (retraining on both old and new data) and regularization to prevent the model from forgetting previously learned speech patterns, thus maintaining the ability to recognize previously encountered speech while improving its performance on newer data.
III. **Monitoring and Evaluation:** Regularly monitor the model’s performance on a validation set, ensuring that the continuous updates lead to meaningful improvements without negatively affecting the model’s ability to recognize existing speech patterns.