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

Self-Supervised Learning (SSL) Pipeline for Dysarthric Speech

Addressing the challenges posed by dysarthric speech, including variability, noise, and limited labeled data, this pipeline focuses on utilizing self-supervised learning techniques for robust performance. The details are as follows:

---

1. Data Pre-Processing

1. **Standardization**:
    - All audio data is standardized to 16 kHz, 16-bit PCM format using tools such as FFMPEG to ensure uniform input quality.
2. **Segmentation**:
    - Voice Activity Detection (VAD) is employed to detect and remove silences exceeding 1 second.
    - Audio is segmented into 20-second chunks for computational efficiency and model compatibility.
3. **Feature Extraction**:
    - 80-dimensional log-Mel spectrograms are extracted using a 25 ms frame window and a 10 ms overlap to capture speech features comprehensively.
4. **Noise Filtering**:
    - An Audio Event Detection (AED) model, based on the Xception-small architecture, filters background noise like music and alarms, retaining only speech-specific segments.
5. **Data Augmentation**:
    - Random cropping is applied to audio segments longer than 5 seconds, creating a diverse training dataset that accounts for various speech patterns.

---

2. Pre-Training Framework

1. **Model Architecture**:
    - An encoder-decoder architecture similar to Lfb2vec is utilized. The encoder comprises 6-layer bidirectional LSTMs or transformer-based models for contextual understanding.
2. **Masked Feature Learning**:
    - Random masking is applied to approximately 6.5% of the input frames. Masked frames are passed through the encoder, which predicts their target representations.
3. **Contrastive Loss**:

- The flatNCE loss function is employed for its computational efficiency and reduced variance compared to InfoNCE. It provides robust learning from both positive and negative samples.

---

3. Fine-Tuning

1. **Supervised Fine-Tuning**:
   - After pre-training, the model is fine-tuned on dysarthric speech datasets to adapt to articulation and pacing nuances. The senone-based acoustic models are leveraged for better phoneme alignment and decoding.
2. **Two-Stage Training**:
   - In the first stage, pre-trained encoder layers are frozen, and new linear projection layers are trained.
   - In the second stage, the entire model is fine-tuned to optimize overall performance.

Continuous Learning Framework

To ensure adaptability to new speakers and evolving speech patterns, a continuous learning framework is integrated into the pipeline:

1. Data Collection and Annotation

- Real-world data is collected via mobile apps or clinical collaborations. Semi-supervised learning is used to reduce dependency on human annotations, leveraging SSL for unlabeled data.

2. Online Learning

- Employ Elastic Weight Consolidation (EWC) to preserve previously learned patterns while incorporating new data.
- Maintain a replay buffer of past data to periodically retrain the model and prevent catastrophic forgetting.

3. Feedback Integration

- User feedback on recognition errors is directly incorporated as pseudo-labels. This ensures the model adapts to user-specific needs over time.

4. Incremental Updates

- Fine-tune upper layers (e.g., linear projections) while keeping core encoder layers fixed, balancing efficiency and adaptability.

5. Federated Learning for Privacy

- Deploy federated learning to allow edge devices to locally adapt models to users while sharing gradients for global updates. This approach preserves user privacy and scales effectively.

Summary

This SSL pipeline is tailored to handle dysarthric speech's unique challenges, emphasizing robust pre-processing, contrastive loss optimization, and adaptive learning. The continuous learning framework ensures the model evolves with user-specific feedback and changing data, providing a scalable and effective solution for real-world deployment.