Question: Propose a model self-supervised learning pipeline to cater dysarthric speech and describe how you would do continuous learning.

Introduction

Dysarthric speech is characterized by slurred or slow speech, caused by neurological damage. Given how variable such speech patterns could be, using the self-supervised learning (SSL) approach from the paper, together with datasets for dysarthric speech (e.g., TORGO, EasyCall corpus), might provide good results.

Data Preprocessing Pipeline

As the paper suggests, a robust preprocessing pipeline could look like this:

- Audio Standardization and Segmentation: Convert audio to 16kHz, 16-bit PCM and using Voice Activity Detection (VAD) to remove silences and segment audio
- Feature Extraction: Generate 80-dimension Log-Mel features
- Noise Reduction and Speech Isolation: Use a Xception-based Audio Event Detection (AED) model to filter background noise and isolate speech
- Augmentation: Distort normal speech to mimic dysarthric features, by adding noise and altering speech pace.

Contrastive Loss on Paired Dysarthric and Normal Speech

Leveraging on the concept of contrastive loss, instead of comparing audio segments from the same distribution (e.g., from dysarthric speech), we could pair data for dysarthric speech and its corresponding normal speech. This would enable the normal to learn features that bridge the gap between dysarthric and normal speech patterns.

Self-Supervised Learning

After the model has been pre-trained using contrastive loss, we could deploy a SSL approach by continuing to learning on unlabeled dysarthric speech data. Furthermore, we could train the model using unlabeled datasets from multiple languages, which could allow the model to adapt to the shared representations across languages (e.g., slurred speech patterns).

Performance Metrics

We could use the following metrics to assess the model's performance:

- Word Error Rate (WER): Measures the proportions of errors, including substitutions, deletions and insertions
- Perceptual Evaluation of Speech Quality (PESQ): Evaluates the quality of speech signals, comparing predicted transcriptions against human intelligibility.

Continuous Learning

To enable continuous learning, we could consider the following approaches:

- Feedback: Allow users to provide feedback (confirm or provide correct the outputs of a model), which is then used to tune the model
- Continued Data Collection: New dysarthric speech samples should be collected to regularly tune the model
- Curriculum Learning: Start with mild dysarthric speech first, which is easier to learn and gradually introduce more severe dysarthric speech
- Domain Adaptation: Fine-tune the model on domain-specific datasets (e.g., for first responders in ambulances, the model could be adapted to audio that is muffled by face masks and occluded by the siren of the ambulance.