Self-Supervised Learning (SSL) pipeline implementation for dysarthric speech (including integration of continuous learning in CI/CD context)

Dysarthric speech presents a major challenge for automatic speech recognition (ASR) due to its significant deviations from normal speech patterns in pronunciation, phoneme elongation, and intelligibility. Additionally, the scarcity of labelled dysarthric speech data limits the effectiveness of traditional supervised models (Solemanpour et al., 2024). Recent advances in self-supervised learning (SSL) for ASR, particularly its application in hybrid ASR (Karimi et al., 2022), suggest that SSL can overcome these challenges.

This essay proposes a custom SSL pipeline to improve ASR for dysarthric speakers, incorporating data pre-processing, SSL pre-training, fine-tuning, and a continuous learning framework to address data drift and feature-based differences (e.g., regional, severity, and type).

Pre-Processing Strategy

Due to the high variability in dysarthric speech, effective data curation is essential. We identify two key transformations: voice activity detection (VAD) to remove excessive silences and pauses, and automatic event detection (AED) using an Xception-based model to filter out background noise, breathing artifacts, and unintended vocalizations. These refinements ensure the model learns from cleaner, more meaningful speech data.

Pre-Training Strategy

We adopt contrastive learning in non-streaming mode to ensure efficient learning of slurred dysarthric utterances. Key modifications include a dysarthria-specific masking strategy, prioritizing prolonged phoneme segments over random masking, and using FlatNCE over InfoNCE for unbiased mutual information estimation. To enhance adaptability, we propose a multi-head multi-severity SSL model with a shared encoder for general dysarthric speech representation and separate projection/decoder layers for distinct severity levels (mild, moderate, severe). This approach ensures fine-tuned severity-specific learning while maintaining generalization across different dysarthric speech types. Unlike supervised pre-training, SSL enables the model to generalize more effectively, a crucial advantage given the high variability in dysarthric speech.

Fine-Tuning Strategy

Fine-tuning is critical for ASR accuracy and robustness. We propose using Latency-Controlled BiLSTM (LC-BLSTM) to handle irregular phoneme durations and a linear projection layer with a language model to improve decoding and transcription accuracy. A two-stage fine-tuning approach is implemented: first, freezing LC-BLSTM layers while training the projection layer to stabilize learned speech representations, followed by full model fine-tuning to adapt to dysarthric speech patterns while minimizing catastrophic forgetting. Additionally, a 5-gram language model (LM) is incorporated for enhanced phoneme correction through context-aware transcription.

Continuous Learning Framework

Given the progressive nature of dysarthria, continuous learning is essential to maintain ASR performance. Our framework includes human-in-the-loop feedback, where confidence scores identify uncertain transcriptions for review, allowing real-time corrections that improve the model. Pseudo-labeling enhances training data by incorporating high-confidence ASR outputs while manually validating low-confidence transcriptions. Additionally, we propose dysarthric speech synthesis (Solemanpour et al., 2024) by integrating a severity level coefficient and pause-insertion model, generating synthetic dysarthric speech for data augmentation. These additional data points can be used to supplement re-training datasets for domain adaptation and continuous learning.

Conclusion

This SSL-based pipeline provides a scalable and adaptable solution for dysarthric speech ASR by integrating targeted pre-processing, contrastive SSL training, fine-tuning, and continuous learning. Future works can explore transformer-based architectures such as Conformers and domain-specific large language models (LLMs) to further optimize dysarthric speech recognition.

References

Karimi, M., Liu, C., Kumatani, K., Qian, Y., Wu, T., & Wu, J. (2022). Deploying self-supervised learning in the wild for hybrid automatic speech recognition. arXiv preprint arXiv:2205.08598.

Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2023). Accurate synthesis of dysarthric speech for ASR data augmentation. arXiv preprint arXiv:2308.08438.