Self-supervised Learning Pipeline for Dysarthric Speech Recognition

Lye En Lih

Dysarthric speech recognition is a challenging problem due to the variability in speech patterns caused by motor impairments. A self-supervised learning (SSL) approach can significantly enhance recognition by leveraging large amounts of unlabelled speech data. This essay outlines an SSL pipeline for dysarthric speech recognition and explores methods for continuous learning to ensure adaptability over time.

Firstly, developing an SSL-based dysarthric speech recognition system requires **data collection and preprocessing**. Since dysarthric speech varies significantly across individuals, gathering a diverse dataset representing different severities and speaker demographics is crucial. Given the limited availability of transcribed dysarthric speech, an SSL approach is ideal for leveraging raw, unlabelled audio. To improve the quality of the dataset, an Audio Event Detection (AED) system can be employed to filter out non-speech segments, reducing noise and ensuring that only relevant speech is included in training. Additionally, segmentation techniques can break long recordings into manageable utterances, facilitating efficient learning.

With a clean and well-structured dataset, the next step is **self-supervised pre-training** using models like wav2vec2, HuBERT, or a customised variant trained specifically for dysarthric speech. These models learn speech representations by predicting masked portions of the input audio, enabling them to capture essential phonetic and acoustic features. However, dysarthric speech often has atypical articulation patterns, necessitating domain-specific adaptations. Instead of training on general speech datasets, the SSL model should be pre-trained on dysarthric speech data. This can be achieved using a contrastive loss function, **such as InfoNCE or flatNCE**, to learn meaningful representations.

After pre-training, **fine-tuning with supervised learning** is necessary to adapt the model for dysarthric speech recognition. A smaller labelled dataset of dysarthric speech is used to optimise the model's output for specific speech-to-text tasks. Furthermore, data augmentation techniques such as time-stretching, pitch-shifting, and noise injection can be used to expand the dataset and improve model generalisation. Then, a transfer learning approach can be applied, where an SSL model pre-trained on standard speech data is fine-tuned on dysarthric speech. Multilingual pre-training with multi-head architectures can also help improve performance, allowing for cross-lingual transfer while maintaining language-specific adaptations.

To ensure long-term adaptability, a **continuous learning framework** can be implemented. Dysarthric speech characteristics may change over time due to disease progression or therapy, requiring regular updates to the model. This can be achieved through incremental learning, where the model is periodically updated with new dysarthric speech data. Pseudo-labelling can generate labels for untranscribed speech, enabling semi-supervised learning without extensive manual annotation. Moreover, domain adaptation techniques can help the model adjust to shifts in speech patterns by continuously refining its representations. Confidence-based learning mechanisms can further enhance performance by prioritising high-confidence predictions in self-training, reducing the risk of propagating errors.

In conclusion, an SSL pipeline tailored for dysarthric speech recognition involves data preprocessing, self-supervised pre-training on domain-specific data, fine-tuning with labelled speech, and implementing continuous learning strategies. By leveraging the strengths of SSL and adapting models to the unique characteristics of dysarthric speech, this approach enhances recognition accuracy, making speech technology more inclusive for individuals with speech impairments.