HTX xData Technical Test (Data Scientist)

Essay Question - Model Self-supervised Learning Pipelines for Dysarthric Speech

Dysarthric speech is a motor speech disorder resulting from neurological impairments that affect the control and coordination of the muscles used for speaking. This condition leads to reduced intelligibility and significant variability in speech patterns, including inconsistent pronunciation, changes in pitch, and prolonged silences. Such speech characteristics make dysarthric speech challenging for conventional Automatic Speech Recognition (ASR) systems, which are typically trained on large, curated datasets of standard speech. Addressing this challenge requires innovative approaches, particularly self-supervised learning (SSL), which excels in leveraging untranscribed and uncurated audio data to learn robust speech representations.

To develop an SSL pipeline tailored for dysarthric speech, it will firstly involve data pre-processing. Dysarthric speech often contains non-speech sounds, background noise, and prolonged silences, which hinder conventional model performance. The pipeline should incorporate Voice Activity Detection (VAD) libraries to remove long silences and segment audio into manageable and meaningful chunks. Additionally, Audio Event Detection (AED) models, such as Xception-based classifiers, can also be used to isolate speech segments from irrelevant noise or non-speech events. Once the audio is pre-processed, features like log-Mel spectrograms can be extracted to effectively represent the speech signal. These steps ensure the data fed into the SSL model is of high quality and representative of the distinct features of dysarthric speech.

A possible SSL pipeline could be an ASR model such as wav2vec2 or HuBERT, fine-tuned to address the unique characteristics of dysarthric speech. Since SSL does not rely on large annotated datasets, it is ideal for scenarios where labeled dysarthric speech is scarce. The model learns speech representations by masking parts of the input and predicting them based on contextual information. For dysarthric speech, in-domain pre-training on a curated dataset of untranscribed dysarthric speech is crucial, and can be sourced from speech banks or collected in collaboration with healthcare organizations. A contrastive learning objective, such as flatNCE, could be used due to its stability and efficiency compared to InfoNCE. This method allows the model to extract meaningful representations despite the irregularities inherent in dysarthric speech.

Effective fine-tuning of the SSL model is crucial for adapting it to dysarthric ASR tasks. Initially, the pre-trained model’s core layers are frozen while training a task-specific projection layer using a smaller labeled dataset. This stage is followed by end-to-end fine-tuning of the entire model to improve its performance further. To enhance decoding accuracy, a language model tailored to dysarthric speech patterns, such as a 5-gram model, can be incorporated into the system.

Continuous learning is essential for maintaining the performance of the ASR system in real-world scenarios. Dysarthric speech varies significantly across individuals, necessitating periodic updates to the model. Online learning techniques, such as teacher-student training, can facilitate this process by generating pseudo-labels for newly collected speech data. These pseudo-labels can then be used to refine the model further. Reinforcement learning, where user feedback is integrated into the training process, provides an additional layer of adaptability, allowing the system to dynamically respond to transcription errors and improve over time.