Self-Supervised Learning Pipeline for Dysarthric Speech [1]

1. **Data preprocessing**

Dysarthric speech is characterised to be slurred, slow, or unclear articulation [2]. Therefore, the first step is to handle the variability and noise within the dysarthric speech data. The preprocessing pipeline mainly includes:

- ➢ Audio conversion of raw audio data into a standardised format (e.g., 16 kHz, 16-bit PCM) using tools like FFMPEG.
- ➢ Removal of long pauses (>1 second) using Voice Activity Detection (VAD), thereafter segment the data into chunks of maximum length of 20 seconds.
- ➢ Extract 80-dim log-Mel features by converting audio segments with a feature processing time window of 25-ms with 10-ms window shift.
- ➢ Lastly, identify between background noise (e.g., music in the background, non-speech audio) from speech audio using Audio Event Detection (AED).

2. **Self-supervised learning (SSL) pre-training**

Next, the SSL model can be trained using the Lfb2vec framework. The main steps for SSL pre-training include:

- ➢ The Lfb2vec framework, similar to wav2vec 2.0, will perform masking of certain parts of the speech (log-Mel features). Random time steps are selected to be masked with a probability of 0.065, and the following 10 time steps are also masked.
- ➢ After masking, the features are passed through a 6-layer Bi-directional LSTM encoder with 600 hidden dimensions. The masked features are then passed through a 20-dimensional linear projection layer, L2-normalised, and output as masked context vectors. The same log-Mel features are passed through another 20-dimensional projection layer to create target vectors, which are also L2-normalized.

- ➢ A contrastive loss function is then used to compare the masked context vectors and target vectors. During training, 100 negative samples are drawn randomly from the same utterance, but from other positions of the target vectors.
- ➢ Both in-domain and out-domain monolingual SSL, as well as out-domain multilingual SSL are then employed. Given the difficulty of collecting 100 negative samples from the same utterance in streaming SSL with short chunk lengths, Lfb2vec is trained using a non-streaming pre-training approach, where the complete log-Mel features of an entire utterance are fed into the model.

Continuous learning

Finally, the model performance will need to be improved over time with different speakers and conditions. Furthermore, the monitoring of the model's performance over time is important to detect potential data drift as new dysarthric speech data is tested. The main steps for continuous learning include:

- ➢ The model would be trained using an online learning setup, where it continuously receives small batches or even single data points from new speech data. The model parameters are updated incrementally after processing each new batch of data (or each individual data point). This enables the model to adapt in real-time as new speech data arrives.
- ➢ Elastic Weight Consolidation (EWC) is then used during the fine-tuning process to avoid catastrophic forgetting while allowing for continuous adaptation. As new speech data arrives, the model is fine-tuned by adjusting its parameters, but EWC ensures that important features learned previously are not drastically altered.

By integrating both online learning and EWC, the model can incrementally adapt to new speech data in real-time, while maintaining stability in previously learned features.

References

[1] M. Karimi, C. Liu, K. Kumatani, Y. Qian, T. Wu, and J. Wu, “Deploying self-supervised learning in the wild for hybrid automatic speech recognition,” May 17, 2022, *arXiv*: arXiv:2205.08598. doi: 10.48550/arXiv.2205.08598.

[2] “Dysarthria - Symptoms and causes,” Mayo Clinic. Accessed: Mar. 25, 2025. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/dysarthria/symptoms-causes/syc-20371994

Done by: Nigel Wee

Words: 511