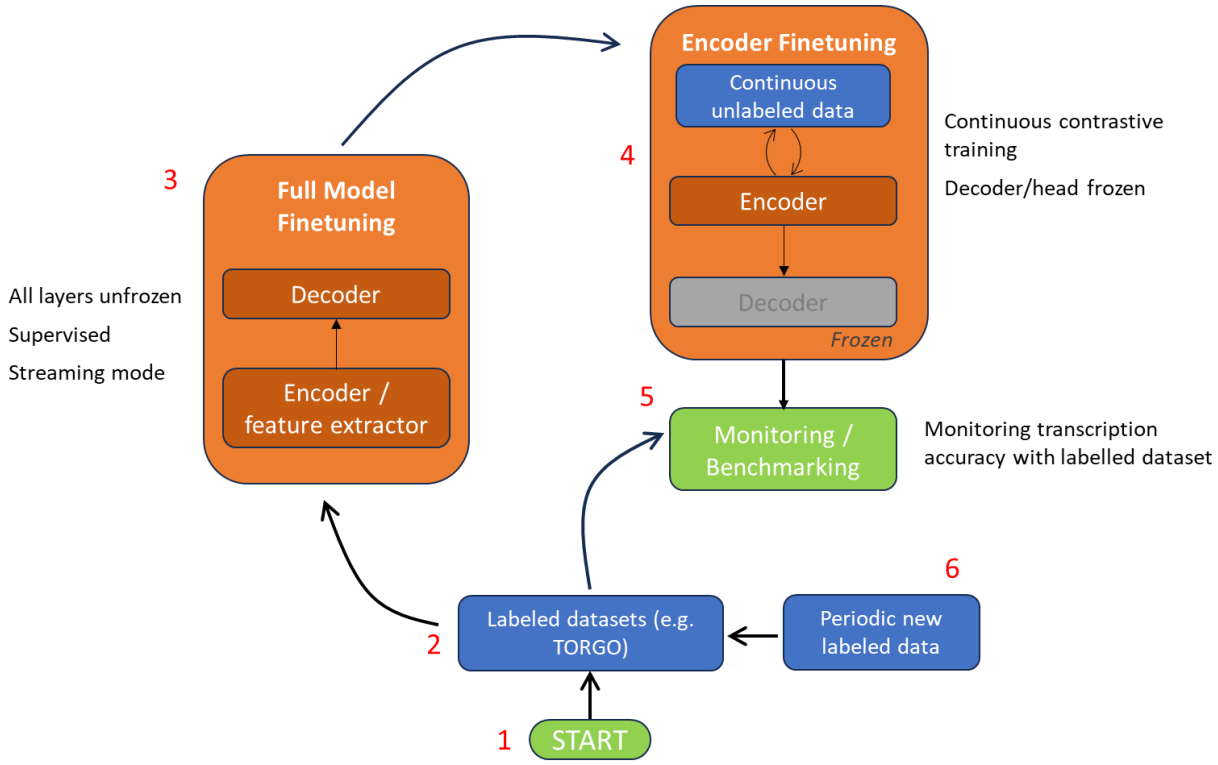
Proposal: Self-Improving Model for Automatic Speech Recognition of Dysarthric Speech

Dysarthric Speech (DS) is a speech disorder caused by lack of coordination due to neurological conditions like cerebral palsy, resulting in slurred speech. In this document, we propose an approach for a Self-Supervised Learning (SSL) pipeline for a continuously learning, Automatic Speech Recognition (ASR) model, aimed at providing real-time transcription. The strategy is developed with reference to findings by [1]. A high-level overview of the pipeline is shown below.



Pipeline Description:

STAGE No.	STAGE	DETAILS
1	Pipeline begins	*N.A.*
2	Labelled data	• Labelled data curated and cleaned for supervised learning in stage **3**.
3	Full model finetuning	• Base model finetuned with all layers unfrozen, including both encoder and decoder/head layers. Encoder finetuned because dysarthric speech patterns may deviate significantly from typical speech in vocalization, articulation, etc. • Considering later deployment as real-time transcribers, streaming mode (elaborated below) is used here.
4	Encoder finetuning	• Exploit higher availability of unlabeled audio data for SSL and encoder-finetuning.

		• Audio data gathered via apps installed on smartphones carried by participants with dysarthria. • Decoder layers frozen. Encoder layer finetuned with contrastive training.
5	Monitoring	• Changes in model performance monitored through frequent evaluation against labelled dataset.
6	New labeled data	• New labeled data collected where feasible (e.g. persons with dysarthria reading pre-written scripts) • Incorporated into labelled dataset and full model finetuning (stage **1**) triggered.

The technical setup is given below, primarily adopting the methodology from [1] as starting parameters/setting. The exact values can be experimented with and continuously integrated.

- Models – a multi-armed bandit approach is suggested, starting both the Wav2Vec2 and Whisper model on the pipeline and selecting the superior model for later deployment/continuous learning.
    - Wav2Vec2 – may be more suitable for finetuning feature extractor layers.
    - Whisper – inherently more robust to non-pristine audio recordings.
- Pre-processing – For audio input data:
    - FFMPEG for audio conversion to 16kHz and 16-bit data
    - Voice Activity Detection (VAD) to filter silences above 1s
    - Audio Event Detection (AED) to filter speech from background noise.
    - Conversion to log-Mel features (25ms window and 10ms window shift)
- Streaming mode – For real-time transcription, training in streaming mode is more appropriate:
    - 20 frame chunks lengths.
    - 20 frame look-ahead ahead length
- Continuous learning:
    - Kafka to manage streaming data for encoder finetuning stage, Airflow to trigger model finetuning processes.
    - Mitigate catastrophic forgetting from long-running training with approaches like learning without forgetting [2]
- Contrastive training:
    - Using flatNCE [3] as contrastive loss
    - 100 negative samples for contrastive training

Bibliography

[1] L. C. K. K. Q. Y. W. T. W. J. Karimi M., "Deploying self-supervised learning in the wild for hybrid automatic speech recognition," *arXiv preprint arXiv:2205.08598.,* 2022.

[2] Z. H. D. Li, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 12, pp. 2935-2947, 2017.

[3] G. Z. L. X. G. Q. C. L. G. S. C. T. X. Y. Chen J., "Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce," *arXiv preprint arXiv:2107.01152,* 2021.