Dysarthric speech is speech that sounds slurred, slow, or hard to understand because the muscles used for talking are weak or don’t move properly. It presents high articulation variability, and often limited transcribed data. This makes self-supervised learning (SSL) ideal.

The first step would be to gather as much speech as possible from people with dysarthria, even if it is not fully transcribed. We clean the recordings by trimming long silences, cutting them into shorter clips, and filtering out background noise. Dysarthric speech often sounds slower or more slurred. Therefore, we create “synthetic” examples by altering speed, loudness and clarity. This gives the model a richer variety of examples to learn from, even before we have detailed labels.

Instead of training only on labelled transcripts, we will use SSL. In SSL, the model tries to predict hidden parts of the speech signal from the surrounding context. This allows it to discover patterns of sounds, timing and articulation without needing full transcripts. By pre-training the model on both typical and dysarthric speech, it learns a general representation of speech but with better coverage of the variations seen in dysarthria.

After this pre-training, we fine-tune the model for the actual task. For example, turning speech into text or classifying the severity of impairment by using the smaller set of recordings that do have transcripts. It needs far fewer labelled examples to reach good performance as the model already understands speech structure from SSL. We can also add small adapter layers that are trained individually for each speaker so that the main model stays stable while still personalising to a person’s unique speaking style.

Speech patterns can change over time, and new speakers may join the system. To keep the model up to date, we can introduce a continuous learning loop. When the model is deployed, it stores a small sample of the new audio it hears. It runs its own prediction and assigns a confidence score. High-confidence predictions can be treated as provisional labels and low-confidence or clearly wrong predictions can be flagged for human correction. This way, we can build up a steadily growing pool of new training data.

At regular intervals, we can retrain the model on a mixture of old and new data. We should always include a small “replay” of earlier examples so that the model does not forget what it learned before. For personalisation, the lightweight adapter layers can be updated much more frequently without touching the main model. We can also monitor performance metrics such as error rate and we will roll back updates if we see a drop.

By combining self-supervised pre-training, targeted fine-tuning and a carefully managed continuous learning process, we get a speech model that adapts well to the challenges of dysarthria and keeps improving as it encounters new speakers and new speech patterns.