Essay question – please read https://arxiv.org/pdf/2205.08598.pdf and propose a model self-supervised learning pipeline to cater dysarthric speech and describe how you would do continuous learning in 500 words.

Dysarthria is a motor speech disorder that affects the physical production of speech due to muscle weakness or paralysis. Self-supervised learning (SSL) offers a promising approach to improve automatic speech recognition (ASR) systems for individuals with dysarthria by leveraging large amounts of unlabeled data. This essay proposes a pipeline for SSL tailored to dysarthric speech and discuss strategies for continuous learning.

The work by Karimi et al. presents a method to utilise uncurated audio data in SSL, from data preprocessing steps to deploying a hybrid ASR model. The preprocesing pipeline discusses the effect of audio event detection (AED), and the choice of optimizer and learning rate scheduling is also analysed. Recently developed contrastive loss functions such as flatNCE are explored as well.

First step would be to gather a diverse dataset of dysarthric speech recordings, ensuring a wide range of severity levels and speakers. As suggested by the referenced work, the audio data can be preprocessed by: (1) using voice activity detection to filter and remove long silences, as well as segment the audio data, (2) converting to log-Mel features, (3) using AED to distinguish speech from other audio events, (4) using AED's prediction to filter the data. Advanced techniques like Gammatone Frequency Cepstral Coefficients (GFCCs) which are more robust to variations in dysarthric speech, can be considered in place of Mel frequency to generate input features.

Using the filtered data, conduct SSL pre-training. Two vector outputs are compared for contrastive loss: a masked context vector which is produced by randomly masking portions of the input and passing it through an encoder, and a target vector that is obtained directly from linear projection and L2 normalisation of the audio features. The trained encoder can be used in a streaming ASR pipeline as an encoder for input features from the streaming audio. This is the masked contrastive learning procedure that the referenced work proposes via Lfb2vec, but there are other SSL tasks such as contrastive predictive coding to predict future frames that can be explored. The encoder architecture used in the work is a 6-layer Bi-LSTM, but other architectures such as Transformers can be utilised for better capturing long-range dependencies in dysarthric speech signals.

After pretraining the SSL model, finetuning is carried out on a smaller labelled dataset of dysarthric speech. Continuous learning can be implemented by periodically updating the model with new data. This can be done using techniques like incremental learning, where the model is updated with new examples without forgetting previous knowledge. The performance of the updated model can be evaluated using standard metrics such as word error rate or phoneme error rate.