Introduction

Dysarthric speech is a challenge in speech processing due to variability in speech patterns caused by motor impairments. Traditional supervised learning models requires large amounts of labeled data which can be difficult and time-consuming to obtain, especially for dysarthric speech. Self-supervised learning (SSL) can help by allowing models to learn useful representations from unlabeled data. This essay proposes a self-supervised learning pipeline for dysarthric speech recognition, incorporating fine-tuning and continuous learning to ensure the model evolves with new data over time.

Model Pretraining and Fine-tuning

The SSL pipeline begins with pretraining on a large corpus of unlabeled speech data. The aim is to learn meaningful speech representations without annotated labels. Contrastive learning is a key technique, where the model distinguishes between similar and dissimilar speech samples. This is achieved by creating positive (e.g., variations of speech from same individual) and negative (e.g., different speakers) audio pairs, training the model to maximize similarity between positive pairs while minimizing similarity for negative pairs. Another effective pretraining strategy is masked prediction, where segments of the audio are randomly masked and the model learns to predict the missing portions. This enhances the model's ability to understand distorted speech patterns such as irregular articulation.

Once pretraining is complete, the model is fine-tuned on a smaller labeled dysarthric speech dataset. Fine-tuning leverages transfer learning, where the model's pre-learned speech representations are adapted to recognize dysarthric speech more accurately. By updating model weights based on labelled data, it improves the model's ability to map learned speech patterns to actual transcriptions, helping it to recognise characteristics like stuttering, slurred pronunciation etc.

Continuous Learning

To maintain effectiveness, we can implement continuous learning where the model is periodically updated with small batches of new labelled speech samples. This allows the system to evolve over time while preserving previously acquired knowledge. To reduce catastrophic forgetting, we propose reintroducing subsets of past data during retraining, ensuring the model's ability to generalize. Additionally, adaptive learning rates help maintain model stability and prevent overfitting to new data while preserving generalization.

Conclusion

The proposed self-supervised learning pipeline for dysarthric speech recognition consists of pretraining, fine-tuning and continuous learning to ensure model remains robust and adaptable. Contrastive learning and masked prediction can improve representational learning, while continuous learning and knowledge preservation techniques ensure the model remains up-to-date. This approach provides a scalable and data-efficient solution for improveing speech recognition in individuals with dysarthria.