Task 6.

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. Your answer can be saved as essay-ssl.pdf under the main repository.

To cater to dysarthric speech recognition using self-supervised learning, this essay proposed using Lfb2Vec model for pre-training and fine-tuning on labelled dysarthric speech datasets such as UASpeech (Zhang et al., 2020) and TORGO (Davis et al., 2007), which contain speech samples with varying degrees of dysarthria, as part of the pipeline. Prior to feeding the dysarthric speech datasets to the model, preprocessing steps which include audio normalization to account for variations in loudness and pitch across dysarthric speech, resampling to 16 kHz to be compatible with Lfb2Vec and feature extraction by converting the resampled audio data into Mel-spectrograms/Log-Mel spectrogram should be carried out. Next, the pre-trained Lfb2Vec is fine-tuned on labelled dysarthric speech datasets such as UASpeech and TORGO by tuning hyperparameters such as learning rate (1e-5 to 1e-3), batch size (16-64), optimizer choice (Adam or AdamW), and regularization (L2 regularization with weight decay between 1e-5 to 1e-3) techniques to avoid overfitting and ensuring generalization capabilities. Feature augmentation techniques such as noise injection, time-stretching and pitch-shifting can also be included during fine-tuning to improve model robustness. Finally, the performance of the fine-tuned model can be evaluated using word error rate to assess its accuracy in transcribing dysarthric speech.

For continuous learning of the fine-tuned model for dysarthric speech recognition, few strategies can be implemented so that the model is updated with new data to adapt to evolving speech patterns. Firstly, the model can be retrained on new data by freezing lower layers and updating only the higher layers so the model can retain previously learned features while adapting to new speech variations. This process is known as incremental learning/fine-tuning. In addition, self-regularization using elastic weight consolidation is one technique to prevent catastrophic forgetting in neural networks (Kirkpatrick et al., 2017). It works by adding a regularization term to the loss function that penalizes large changes in important parameters, thereby allowing the model to learn new patterns while preserving the knowledge gained from earlier training. In addition, Wav2Vec 2.0, a self-supervised learning model, has demonstrated success in fine-tuning speech data with limited annotated examples (Baevski et al., 2020), thereby making it a suitable model for incremental fine-tuning in dysarthric ASR. Furthermore, the model should regularly be updated with new data from real-world users so different types or severities of dysarthria can be incorporated. This strategy is known as domain adaptation.

Lastly, continuous learning requires ongoing evaluation using metrics such as word error rate and character error rate to ensure that the model adapts well to the new data without compromising on the model performance on previously learned speech patterns.

(428 words)

References

Baevski, A., Zhou, H., & Mohamed, A. R. (2020). *Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. In Proceedings of NeurIPS 2020.

Davis, C., et al. (2007). *TORGO: A Dysarthric Speech Corpus for Recognition and Synthesis Research*. INTERSPEECH 2007.

Kirkpatrick, J., et al. (2017). *Overcoming Catastrophic Forgetting in Neural Networks*. Proceedings of NeurIPS 2017.

Zhang, Z., et al. (2020). *UASpeech: A Speech Corpus for Dysarthric Speech Recognition*. Proc. Interspeech, 2020, 492-496.