Abstract:

The recognition of accented speech still remains a dominant problem in Automatic Speech Recognition (ASR) systems. We approach the classification of accented English speech through the Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) architecture which has been shown to perform well on a variety of speech tasks. Three models are proposed: one trained from scratch, another two models (one using data augmentation and a baseline model) fine-tuned from the checkpoints of speechbrain/spkrec-ecapa-voxceleb (VoxCeleb). Our results show that the model fine-tuned with data augmentation yield the best results. Most of the misclassifications were structured and expected due to accent similarities, such as the American and Canadian accents. We also explored the internal categorization of embeddings through t-SNE, a dimensionality reduction technique, and found that there was a level of clustering based on phonological similarity. For future work, we would like to explore the implementation of this accent classification system in our suggested framework to improve ASR performance by making it more inclusive to accented speech.