Abstract

Despite recent improvements in speaker-independent automatic speech recognition (ASR), the performance of large-scale speech recognition systems is still significantly worse on dysarthric speech than on standard speech. Both the inherent noise of dysarthric speech and the lack of large datasets add to the difficulty of solving this problem. This thesis explores different approaches to improving the performance of Deep Learning ASR systems on dysarthric speech.

The primary goal was to find out whether a model trained on thousands of hours of standard speech could successfully be fine-tuned to dysarthric speech. Deep Speech – an open-source Deep Learning based speech recognition system developed by Mozilla – was used as the baseline model. The UASpeech dataset, composed of utterances from 15 speakers with cerebral palsy, was used as the source of dysarthric speech.

In addition to investigating fine-tuning, layer freezing, data augmentation and re-initialization were also investigated. Data augmentation took the form of time and frequency masking, while layer freezing consisted of fixing the first three feature extraction layers of Deep Speech during fine-tuning. Re-initialization was achieved by randomly initializing the weights of Deep Speech and training from scratch. A separate encoder-decoder recurrent neural network consisting of far fewer parameters was also trained from scratch.

The Deep Speech acoustic model obtained a word error rate (WER) of 141.53% on the UASpeech test set of commands, digits, the radio alphabet, common words, and uncommon words. Once fine-tuned to dysarthric speech, a WER of 70.30% was achieved, thus demonstrating the ability of fine-tuning to improve upon the performance of a model initially trained on standard speech.

While fine-tuning lead to a substantial improvement in performance, the benefit of data augmentation was far more subtle, improving on the fine-tuned model by a mere 1.31%. Freezing the first three layers of Deep Speech and fine-tuning the remaining layers was slightly detrimental, increasing the WER by 0.89%. Finally, both re-initialization of Deep Speech's weights and the encoder-decoder model generated highly inaccurate predictions. The best performing model was Deep Speech fine-tuned to augmented dysarthric speech, which achieved a WER of 60.72% with the inclusion of a language model.