Partially Speaker-Dependent Automatic Speech Recognition Using Deep Neural Networks

Li, Christopher (School: Spring Valley High School)

As people continue to streamline everyday life with technology, the use of speech recognition has been made available in smartphones, speakers, cars, and more. The process of converting audio to text is known as automatic speech recognition (ASR), and current research is focused on improving robustness to speaker variability and background noises. One way to do so is speaker adaptation, which uses a small dataset of speech from one speaker to boost the program's accuracy on the chosen speaker. Due to the literature gap around an accuracy improvement goal for speaker adaptation, this research project aimed to set this goal by directly training ASR programs on the target speaker. It was hypothesized that training the model on data from a specific speaker would improve the model's accuracy when transcribing different speech from the same speaker. Using the Kaldi Toolkit and TIMIT Speech Corpus, speaker-independent and partially speaker-dependent models were trained. Using phone error rate (PER) as a metric for accuracy, it was found that training the model on the target speaker improved mean target speaker accuracy by an absolute PER of 7.4% and mean overall accuracy by 0.5%. Inferential statistics with t-tests revealed that both the increase in target speaker accuracy $t(47.52) = 19.90$, $p < 0.001$ and the increase in overall accuracy $t(57.97) = 3.33$, $p = 0.0015$ were significant. As a result, this experiment presents a successful partially speaker-dependent system that can be used as a goal for novel speaker adaptation approaches.