Emotion Recognition from Human Speech Using Temporal Information and Deep Learning

Kim, John (School: Menlo School)

Emotion recognition from human voice by machine is a challenging task, but it has great potential to make empathic human-machine communications possible. In conventional approaches, extensive studies have been devoted to developing good feature representations, but relatively little effort was made to make use of important temporal information. The goal of this research is to develop a model combining features known to be useful for emotion recognition and deep neural networks to exploit temporal information when recognizing emotion status. A novel model is proposed that combines features known to be useful for emotion recognition and a Deep Neural Network to model the unknown mechanism in recognizing emotion status from the temporal sequence of feature vectors, which consists of Convolutional Neural Networks for both local and global convolution and Long Short-Term Memory layers. Two different model structures are developed and optimized in order to increase recognition rate; finally, the real-life practicality of this system is explored by investigating the effect of different speaker-dependent modes, compared to speaker-independent mode. The performance evaluation is performed on the Berlin Emotional Speech Database. The database consists of 535 utterances from 10 talker with seven different emotions. A benchmark evaluation demonstrates that the proposed model achieves 88.9% recognition rate, replacing the state-of-the-art performance of 86%. Deep analysis using t-Distributed Stochastic Neighbor Embedding validates that the emotion space constructed by the model is similar to that of human perception. However, the recognition rate is degraded by 7% when the model was evaluated in speaker-independent mode, suggesting further research for practical application.