Name(s)
John Kim

Project Number
38068

Project Title
Emotion Recognition from Human Speech Using Temporal Information and Deep Learning

Abstract

Objectives/Goals
Emotion recognition from voice by machine is a challenging task, but it has great potential to make empathic human-machine communications possible. In conventional approaches that consist of feature extraction and classifier stages, extensive studies have been devoted to developing good feature representations, but relatively little effort was made to make proper use of the important temporal information in these features. 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.

Methods/Materials
I propose a model that combines a feature extraction stage known to be useful for emotion recognition and a Deep Neural Network (DNN) to model the unknown mechanism in recognizing emotion status from the temporal sequence of feature vectors, where the DNN consists of Convolutional Neural Networks for local and global convolution and Long Short-Term Memory layers. Two different model structures are developed and compared to conventional approaches. Considering the practical use of emotion recognition systems, the performance gap issue between speaker-dependent/independent modes is addressed. The performance evaluation is performed on the Berlin Emotional Speech Database, which is one of the most widely used databases. The database consists of 535 utterances from 10 talkers, each utterance representing one of seven different emotions.

Results
A benchmark evaluation demonstrates that the proposed model achieves 88.9% recognition rate, replacing the state-of-the-art performance of 86% with a big margin equivalent to 20.7% error reduction rate. A deeper analysis validates that the emotion space formed by internal representations of the proposed model is similar to that of human perception. The recognition rate is degraded by 7%, when the model was evaluated in speaker-independent mode.

Conclusions/Discussion
A novel model is proposed to recognize emotion from human speech. The model consists of acoustic speech analysis to maximize the efficiency of well-known feature extraction and DNN to learn an unknown mechanism of temporal information. The proposed model replaces the state-of-the-art performance with a big margin. Further works include using an extended feature set, evaluation on different databases, and performance improvement in speaker-independent mode.

Summary Statement
A novel DNN model is proposed to recognize emotion from human speech, and achieves 88.9% recognition rate, replacing the state-of-the-art performance of 86% with a big margin equivalent to 20.7% error reduction rate.

Help Received
Dr. Rif Saurous guided me on my research, specifically on the problems of cross-validation.