A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering

ABSTRACT

Speech Emotion Recognition (SER) is a common aspect of human-computer interaction and has significant applications in fields such as healthcare, education, and elder care. Although researchers have made progress in speech emotion feature extraction and model identification, they have struggled to create an SER system with satisfactory recognition accuracy. To address this issue, we proposed a novel algorithm called F-Emotion to select speech emotion features and established a parallel deep learning model to recognize different types of emotions. We first extracted the emotion features from speech and calculated the F-Emotion value for each feature. These values were then used to determine the combination of speech emotion features that was optimal for speech emotion recognition. Next, a parallel deep learning model was established with the speech emotion feature combination as input to train and test for each type of emotion. Finally, decision fusion was applied to the parallel output results to obtain an overall recognition result. These analyses were conducted on two datasets, RAVDESS and EMO-DB, with the accuracy of speech emotion recognition reaching 82.3% and 88.8%, respectively. The results demonstrate that the F-Emotion algorithm can effectively analyze the correspondence between speech emotion features and emotion types. The MFCC feature best describes emotions of neutrality, happiness, fear, and surprise, and Mel best describes emotions of anger and sadness. The parallel deep learning model mechanism can improve the accuracy of speech emotion recognition.