Doctoral Dissertation

Speech emotion recognition based on human auditory characteristics using deep learning methods

Zhichao Peng

Supervisor: Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

September 2020

Abstract

The coming era of the Internet of Everything provides huge development opportunities for the field of human-robot interaction. Speech is the most natural and convenient way for communication between humans and robots. Emotion information from speech can effectively help robots understand the speaker's intentions in natural human-robot interaction. Therefore, speech emotion recognition (SER) is one of the hotspots in current research, which can play an essential role in all human-robot interaction scenarios such as education, medical care, service, etc.

Identifying emotions from speech requires to extract discriminative and robust features that can effectively represent the emotion of speech. However, the traditional acoustic features have problems with weak emotional discrimination and poor noise robustness. The human auditory system can easily perceive the emotional states of speech even in a noisy environment, so this study is to explore auditory representations of computational auditory models and deep learning methods to improve the performance of categorical and dimensional emotion recognition.

Due to the complexity of the human auditory system, the process of speech signal processing is not completely clear, nor which the auditory model can better simulate the human auditory system. Recent psychoacoustic experiments show that temporal modulation cues play an important role in speech perception and contain multi-dimensional spectral-temporal information. Therefore, this study proposes a 3D convolutional neural network (3D CNN) architecture for categorical emotion recognition. In this architecture, 3D CNN is used to extract the discriminative auditory representations from temporal modulation cues by joint spectral-temporal feature learning. The experimental results show that the joint spectral-temporal auditory representations can be extracted using 3D CNN from temporal modulation cues. The results demonstrate that the performance of emotion recognition based on joint spectral-temporal representation can exceed the recognition accuracy compared to that of the state-of-the-art methods.

The high-level auditory representation sequence extracted from 3D CNN is segmented into non-overlapping subsequences, and then LSTM is used to capture the segment-level temporal dependence of subsequences in the previous study. These discontinuous segment-level features cannot fully reflect the dynamic changes of emotions. In addition, existing studies on the attention model only focus on the salient regions of emotion but ignore the continuity of cognition. Inspired by cognitive behavior, this study proposes an attention-based sliding recurrent neural network (ASRNN) to effectively model auditory representation sequence by mimicking the

auditory attention to capture salient emotion regions. In the ASRNN model, a high-level feature representation is obtained continuously through a sliding window, and then a temporal attention model is used to capture salient regions of emotion representation. Moreover, a subjective evaluation experiment is designed to analyze the correlation between the temporal attention model and human auditory attention. The results of the experiments showed that this model could effectively obtain emotional information by capturing salient emotion regions using the ASRNN model. The subjective evaluation shows that the temporal attention model is basically consistent with human auditory attention in recognizing emotions.

In categorical emotion recognition, the 3D convolution is used to extract high-level auditory representation from temporal modulation cues. However, the high-dimensional data space through auditory and modulation filtering is not suitable for dimension emotion recognition. Neuroscience research shows that the cortical encoding of natural sounds entails the formation of multiple representations with different spectral and temporal resolution. Inspired by neuroscience, this study proposes a novel auditory feature, namely multi-resolution modulation-filtered cochleagram (MMCG), to capture temporal and contextual modulation cues. Considering that each modulation-filtered cochleagram in MMCG contains different temporal and contextual modulation cues, a parallel LSTM network structure is designed to model multi-temporal dependencies of MMCG and track the temporal dynamics of speech signal sequence for dimensional emotion recognition. Experimental results show that the MMCG feature can significantly improve the performance of emotion recognition compared with all evaluated features. The results also show that the parallel LSTM can track the temporal dynamics of emotion from each modulation-filtered cochleagram at different scales.

In conclusion, this dissertation investigates different auditory features and some deep learning models for categorical or dimensional emotion recognition according to different features. This study proposes 3D CNN architecture to learn joint spectral-temporal auditory representation from the temporal modulation cues and ASRNN model to capture the salient regions of emotion continuously. Experiment results proved that the proposed methods could effectively extract distinguishable spectral-temporal representations and capture the salient regions from the representation sequence. In addition, this study also proposes the MMCG feature to capture the temporal and contextual modulation cues in different resolutions, and develops a parallel LSTM to capture the temporal dynamics of the MMCG features for dimensional emotion recognition. Experiment results further prove that the proposed methods could effectively capture the temporal dynamics of emotion. The results show that the

proposed deep learning models based on human auditory characteristics have achieved good performance in speech emotion recognition.

Keyword: speech emotion recognition, human auditory characteristics, multi-resolution modulation-filtered cochleagram, 3D convolution, attention-based sliding recurrent neural network