氏　　名	PENG, Zhichao			
学位の種類	博士(情報科学)			
学位記番号	博情第 437 号			
学位授与年月日	令和 2 年 9 月 24 日			
論文題目	Speech emotion recognition based on human auditory characteristics using deep learning methods			
論文審査委員	主査	赤木　正人	北陸先端科学技術大学院大学	教授
		党　　建武	同	教授
		鵜木　祐史	同	教授
		岡田　将吾	同	准教授
		WANG, Longbiao	天津大学	教授

論文の内容の要旨

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

論文審査の結果の要旨

本論文は，ヒトの聴覚特性を模擬した音声信号処理手法の提案，処理結果を入力とした深層ネットワークを用いた感情知覚モデルの構築，および，本モデルの音声に含まれる感情自動推定への適用に関する研究報告である。音声から感情を識別するには，音声中の感情を効果的に表すことができる頑健な特徴を抽出する必要がある。従来の特徴は，感情を区別するには弱く，雑音への頑健性が低いという問題があった。一方人間の聴覚システムは、騒々しい環境でも音声の感情的な状態を簡単に知覚できる。本研究では、深層学習を使用して，ヒトの聴覚機能にもとづいた感情計算モデルを探索し感情認識の性能を向上させた。

本研究では，段階的に 3 種類の手法を提案している。（1）Auditory-based method：蝸牛フィルタと変調フィルタに基づく音声感情認識方法を調査し三次元畳み込みニューラルネットワーク（3D CNN）アーキテクチャを提案した。（2）Attention-based method：Attention に基づくスライディングリカレントニューラルネットワーク（ASRNN）を提案した。（3）Multi-resolution modulation filtered based method：時間的および文脈的変調の手がかりを得るためにマルチ解像度変調フィルタによる Cochleagram（MMCG）を提案した。

研究の結果，手法（1）については，3D CNN アーキテクチャにより，スペクトル時間表現にもとづく範疇的感情認識のための聴覚表現を抽出した。実験結果は，最新競合手法と比較して，提案手法の認識精度が上回る可能性があることを示した。手法（2）については，ヒトの認知行動にもとづいて提案した ASRNN モデルにより，高レベルの特徴表現を継続的に選定し，有意な感情領域を把握した。認識実験結果は，ASRNN モデルを使用して顕著な感情領域を選定することで感情情報を効果的に取得できることを示した。手法（3）については，MMCG に異なる時間的およびコンテキスト変調手がかりが含まれていることを考慮して，並列 LSTM ネットワーク構造により次元的感情認識のための各次元情報の時間的変化を追跡する認識器を設計した。実験結果は，並列 LSTM は各変調フィルタ処理された蝸牛からの感情の時間ダイナミクスを異なるスケールで追跡できており，他の競合手法と比較して MMCG 機能が次元的感情認識の性能を大幅に改善できていることを示した。

以上のように，本研究は新しい概念のもとで，話された言語によらず音声中の感情を高精度で推定する手法を実現したものであり，学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。