学位論文の骨子（情報科学）　　令和 2 年 6 月 18 日

Dissertation Outline (Information Science)

氏名 Name	Zhichao Peng	学生番号 Student Number	S1620432

主指導教員 Supervisor	Masato Akagi	印 Seal	副指導教員 Second Supervisor	Masashi Unoki	印 Seal

☑ 副テーマ指導教員 Advisor for Minor Research Project ☐ インターンシップ指導教員 Advisor for Internship	Mineo Kaneko	印 Seal

＜博士論文題目（仮）＞ (Tentative) Title of Doctoral Dissertation

Speech emotion recognition based on human auditory mechanism using convolutional and recurrent neural networks

＜研究の目的と効果＞　Research Aim and Impact

Speech emotion recognition (SER) plays an important role in robots or virtual agents understanding a speaker’s intentions in natural human-robot interaction (HRI). In SER, one of the central research issues is how to extract the emotion-salient and noise-robust features from the observed speech signals. Most of the current studies mainly focus on traditional hand-tuned acoustic features such as prosodic features, voice quality features and spectral features to find the salient features relevant to emotional speech. When the traditional acoustic features are used for emotion recognition, the recognition performance decreases sharply with the decrease of signal-to-noise ratio (SNR). Therefore, considering how to extract noise-robust features is an important part of SER systems.

As the human auditory system is powerful in processing time-frequency signals and extracting noise-robust features, research has focused on auditory-based speech signal processing for emotion recognition by mimicking the function of the auditory system. An auditory model is considered a robust front-end in SER systems to model the auditory system from the cochlea through the thalamus. The auditory filterbank (AFB) in the auditory model is used as a simulation of the acoustic-frequency analyzer in the cochlea which is responsible to decomposes speech signals into acoustic-frequency components. Furthermore, temporal-envelope extraction from the acoustic-frequency components is modeled to effectively simulate the mechanic-to-neural signal transduction in the inner hair cell (IHC). The modulation filterbank (MFB) is introduced to generate high-resolution temporal-modulation cues provided by the temporal envelope and its modulation-frequency components. Furthermore, psychoacoustic experiments showed that temporal-modulation cues are important for speech perception and understanding. Some studies used the modulation-spectral features (MSFs) for SER. The MSFs are obtained by calculating the spectral centroid, spectral flatness, mean, skewness, kurtosis, and other statistical features from temporal-modulation cues. To reduce the computation of spectral features, however, MSFs are only calculated in each modulation channel and produce time-averaged spectral features. In fact, the 3D spectral-temporal representation of speech is formed from temporal-modulation cues including acoustic frequency components, modulation frequency components, and temporal features after using signal processing of auditory front-ends. This study tries to extract the spectral and temporal distinguished information simultaneously from 3D spectral-temporal representation.

Recently, deep learning has become the best way to find the distinguished features. There are mainly two kinds of research methods, which are used to recognize the emotion in speech based on deep learning. Convolutional neural networks (CNNs) can extract high-level local feature representations using the receptive field of the neuron and have been used for acoustic modeling and feature extraction in SER systems. Recurrent neural networks (RNNs) including long short-term memory (LSTM) are designed to handle long-range temporal dependencies and can be turned into convolutional and recurrent neural networks (CRNN) by combining them with the output of the CNN layer to handle time sequence dependence. Therefore, another research issue is how to extract high-level features from 3D spectral-temporal representation using deep neural networks.

This study aims to propose a robust SER system from speech combining auditory model and deep neural networks in categorical and dimensional representation space. As the human auditory system is powerful in time-frequency signal analysis and processing, an auditory model, which mimics the function of the human auditory system, is used as a front-end to extract spectral-temporal features in SER systems. Additionally, compared with MSFs, these 3D features contain rich spectral-temporal representations and can avoid the modulation-correlation problem. Hence, this study firstly proposed an end-to-end utterance-level SER system using 3D CRNNs based on spectral-temporal representations from an auditory model. 3D convolutional layer is employed to extract frame-level features from 3D spectral-temporal information and then the recurrent layer (LSTM) is employed to obtain temporal-dynamics information in each utterance.

Additionally, in many cases, only a few words in the utterance are emotional, while the majority of the rest are emotionless. Attention model is also introduced to time sequence task with neural networks to find the main emotion part of speech. This study further proposed a SER system that uses attention-based sliding recurrent networks (ASRNNs) to achieve time-sequence classification. This study first use a sliding recurrent network (SRNNs) to continuously extract segment-level internal representations in a sliding-window manner. This study then use a temporal attention model to capture the important information related to emotion from the sequence of internal representations. Finally, the utterance-level representation is formed for SER by applying attention weighs to segment-level representation.

The main contributions of this study are as follows: 1) Proposing a features extraction method to obtain the spectral and temporal distinguished information simultaneously from spectral-temporal representation. 2) Proposing a SER system that uses ASRNNs in a two-step manner. The first step is to use SRNNs to obtain continuous segment-level emotion features, and the second step is to use an attention model to focus on the emotional parts for utterance-level SER. 3) Listening tests show that the attention patterns of the attention model are consistent with that from human.

Experiments results indicate that the proposed system in categorical space can achieve higher unweighted accuracy as well as can obtain better concordance correlation coefficient (CCC) in dimensional space compared with state-of-the-art end-to-end systems. Additionally, compared with the prevalent a SER system, the proposed system has more robust recognition ability in noise environment. Therefore, the proposed system can effectively help robots understand speaker’s intentions in natural human-robot interaction.

＜研究の概要＞ Research Outline

To achieve the research purpose of a SER from speech using a deep neural network model based on human auditory mechanism, the following three major issues are investigated: (1) How to extract the emotion-salient and noise-robust features from speech signals; (2) How to extract high-level features from 3D spectral-temporal representation; (3) How to design deep neural model for dimensional emotion recognition (DER); and (4) Noise-Robustness analysis in categorical and dimensional emotion space respectively.

The first important issue in the design of the SER system is to extract the emotion-salient and noise-robust features. The recognition performance using the traditional acoustic features decreases sharply with the decrease of SNR. Besides, MSFs are inspired from auditory mechanism and robust in noise environment, but these features lose rich spectral-temporal representations using time-average way in each channel. The human auditory system has far superior emotion recognition abilities compared with recent a SER systems, so research has focused on designing a SER systems by mimicking the human auditory system. Additionally, psychoacoustic and physiological studies indicate that the human auditory system decomposes speech signals into acoustic and modulation frequency components, and further extracts temporal modulation cues. Speech emotional states are perceived from temporal modulation cues using the spectral and temporal receptive field of the neuron. This study analyze the characteristics of different AFB and MFB, and tried to find that each emotion has different local or global characteristics in the spectral-temporal representation. Additionally, 3D temporal modulation cues are suitable to extract high-level spectral-temporal representations for convolutional networks.

In terms of the second issue, the main focus is how to extract high-level features from 3D spectral-temporal representation for categorical emotion recognition (CER). This study try to design different methods to extract salient features. First, this study proposed a (CER) system in two-stage way based on auditory model. The first stage is to extract the segment-level robust and compact features from raw audio using multichannel parallel convolutional and recurrent neural networks (MPCRNNs) model. The second stage is to extract the utterance-level statistical features from the different segments belonged to the same utterance, and feed into a SVM classifier to determine the emotional state of the whole utterance. However, the temporal-modulation cues are not considered with this method and not in an end-to-end manner. Second, this study proposed an end-to-end utterance-level SER system using CRNNs based on spectral-temporal representations from an auditory model. The convolutional layer is used to extract high-level multiscale spectral-temporal representations. These representations are different time sequences of varied length, which can be the input of another convolutional layer, or finally the input of a recurrent layer (LSTM) that models the temporal relations in the data. This study attempt to mimic the strong recognition abilities by 3D CNN, and remember the short-term and long-term information using LSTM cells. The proposed method was verified on the IEMOCAP database. The results show that our proposed method can exceed the recognition accuracy compared to that of the state-of-the-art systems. Comparing with the two-stage way, it is similar to the human in recognition ways and is worth to study as an important research. Third, in conventional algorithms, all frames in the utterance are mapped to the same label. Since the labels are annotated for utterances, not for frames, it does not mean all the frames in the same utterance should be mapped to the same label. This study further proposed a SER system that uses attention-based sliding recurrent networks (ASRNNs) to achieve time-sequence classification. This study first use a sliding recurrent network (SRNNs) to continuously extract segment-level internal representations in a sliding-window manner. This study then use a temporal attention model to capture the important information related to emotion from the sequence of internal representations. Finally, the utterance-level representation is formed for SER by applying attention weighs to segment-level representation. To explore the relevance between attention weights and human attention, this study investigated the temporal attention of the human auditory system and compared it with that of an attention model. The results indicate that the entire database has basically the same consistency in terms of the correlation coefficient.

The third important issue is how to design deep neural model for DER. Previous study mainly focuses on CER. Categorical emotion is widely used in people's daily life of several emotions, also known as basic emotions. At present, the six basic emotions proposed by Ekman are widely used in the field of emotion-related research: happiness, anger, sadness, surprise, fear and disgust. In addition to recognizing categorical emotion, people use dimensional emotion to describe more abundant emotion. Dimensional emotion uses continuous numerical values to describe emotional states. It regards emotional state as a point in multi-dimensional emotional space, and each dimension corresponds to different psychological attributes of emotion. The most commonly used dimension affective model is the Valence-Arousal (V-A) model. For speech synthesis with different emotional styles in the V-A dimensional space, which acoustic features are related to which dimensions needs to be researched. To evaluate the agreement level between the predictions of the network and the gold-standard derived from the annotations, the CCC has recently been proposed. In DER, speech signals are allocated a valence and arousal values every short duration (For example: 40 ms in RECOLA database). Its objective function and training method are different from CER. Therefore, a new CNN-LSTM model must be designed. This project plans to use the auditory model as the front-end, and propose a new CNN-LSTM network model, which combines the CCC and mean square error (MSE) as the objective function for DER. This study explore various aspects to improve the prediction performance including: the dominant modalities for arousal and valence prediction.

The last issue is to analyze noise-robustness of SER system based on Auditory Model. On the basis of the above research, the auditory model is further analyzed. Because white noise and environmental noise inevitably exist in the process of human-computer interaction in natural environment, this study try to propose a robust emotional recognition system with noise in categorical and dimensional emotion space respectively. And then compare the traditional emotional recognition system under different white noise or environmental noise SNR with the proposed model. Additionally, considering the influence of gender and other factors on emotional recognition, multi-task learning (emotional recognition as the main task, gender and other recognition as auxiliary tasks) is proposed to further enhance its robustness.

In order to assess the performance of SER, comparison evaluation of performance by the traditional features and the human auditory features for SER is carried out. Meanwhile comparison evaluation of performance by the traditional DNN methods and the proposed CRNN methods is also carried out. In addition, what kind of features should be learned is analyzed, and how well the proposed methods work on mimicking human processing of emotion perception is evaluated.

＜研究の概要（つづき）＞ (continued from previous page)

Our experiments demonstrate that the proposed system can obtain spectral-temporal representations and exhibit better recognition performance compared to that of state-of-the-art SER systems. Experiments results also indicate that the proposed system has more robust recognition ability in noise environment comparing with the prevalent SER system. Moreover, the results of the listening tests indicate that there is strong correlation between human temporal attention and an attention model in CER.

In summary, an auditory model as a front-end can extract rich spectral-temporal information, and the proposed systems in categorical and dimensional emotional space can effectively extract high-level features for SER. This system may be applied to other audio-event recognition, such as speaker recognition and speech recognition.

＜論文の構成案＞ Dissertation Structure

1. Introduction
    - 1.1 Research background
    - 1.2 Research originality
    - 1.3 Research approach
    - 1.4 Dissertation organization
2. Literature review
    - 2.1 Representation of emotion
        - 2.1.1Categorical approach
        - 2.1.2Dimensional approach
    - 2.2 Emotional speech corpus
    - 2.3 Speech emotion recognition
        - 2.3.1 Feature extraction
        - 2.3.2 Classification
    - 2.4 Deep learning
        - 2.4.1 Convolutional neural networks
        - 2.4.2 Recurrent neural networks
3. Concept and proposed scheme
    - 3.1 Human auditory system
    - 3.2 Auditory modelling
        - 3.2.1 Auditory Filterbank
        - 3.2.2 Hair cell transduction
        - 3.2.3 Modulation Filterbank
    - 3.3 Proposed scheme
4. Categorical emotion recognition
    - 4.1 Emotion recognition in two-stage way
    - 4.2 End-to-end emotion recognition
    - 4.3 Attention model
    - 4.4 Listening test for temporal attention
    - 4.5 Noise-robustness analysis
    - 4.6 Discussion
5. Dimensional emotion recognition
    - 5.1 Evaluation measures
    - 5.2 Effectiveness of emotion representation approach
    - 5.3 Experiment
    - 5.4 Results
    - 5.5 Noise-robustness analysis
    - 5.6 Discussion
6. Discussion of the categorical and dimensional perceptions of emotion
    - 6.1 Categorical perception of emotion
    - 6.2 Dimensional perception of emotion
7. Conclusion and future work
    - 7.1 Summary
    - 7.2 Contribution
    - 7.3 Future work

＜研究業績＞ Publication List

1 Journal

[1].○ Z. Peng, Q. Hu, and J. Dang, “Multi-kernel SVM based depression recognition using social media data,” Int. J. Mach. Learn. Cybern., pp. 1–15, 2017.

[2].○ Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Auditory Model as Front-ends for Speech Emotion Recognition Using 3D Convolution and Attention-based Sliding Recurrent Network,” Journal paper (under review).

2 International Conferences

[1].○ Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on gammatone auditory filterbank,” in Proc. 9th Asia-Pacific Signal Inf. Process. Assoc. Annu.Summit and Conf., 2017, pp. 1750–1755.

[2].○ Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral- temporal representation,” In: 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018. p. 1-6.

[3].○ S. Peng, Q. Hu, J. Dang, and Z. Peng. "Stochastic Sequential Minimal Optimization for Large-Scale Linear SVM." International Conference on Neural Information Processing. Springer, Cham, 2017, pp. 279-288

3 Domestic Conferences

[1]. ○ Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “End-to-end speech emotion recognition using 3-d convolutional recurrent neural networks based on modulation spectral features,” Acoustic Society of Japan, Spring, 2018, 2-Q-10.

[2]. ○ Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral-temporal modulation” in JAIST world conference (JWC), 2018

4 Awards

[1]. ○ Fall 2016 Monbukagakusho Honors Scholarship

[2]. ○ Research Grants for JAIST Students 2018-2019.

＜現在の単位修得状況＞ Courses I have obtained credits

	発展科目 Intermediate	先端科目 Advanced	その他 Others	合計 Total	必修 B 科目(S50x) Required Courses
科目数 Number of courses	3	3	1	7	☑ S503 ☐ S501 / S502
単位数 Number of credits	6	6	2	14	