Speech Emotion Recognition Using Multichannel Parallel Convolutional Recurrent Neural Networks based on Gammatone Auditory Filterbank

Zhichao Peng[*, †], Zhi Zhu[*], Masashi Unoki[*], Jianwu Dang[*, †], Masato Akagi[*]
[*]School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[†]School of Computer Science and Technology, Tianjin University, Tianjin, China
E-mail: {zcpeng, zhuzhi, unoki, jdang, akagi} @jaist.ac.jp

Abstract—**Speech Emotion Recognition (SER) using deep learning methods based on computational auditory models of human auditory system is a new way to identify emotional state. In this paper, we propose to utilize multichannel parallel convolutional recurrent neural networks (MPCRNN) to extract salient features based on Gammatone auditory filterbank from raw waveform and reveal that this method is effective for speech emotion recognition. We first divide the speech signal into segments, and then get multichannel data using Gammatone auditory filterbank, which is used as a first stage before applying MPCRNN to get the most relevant features for emotion recognition from speech. We subsequently obtain emotion state probability distribution for each speech segment. Eventually, utterance-level features are constructed from segment-level probability distributions and fed into support vector machine (SVM) to identify the emotions. According to the experimental results, speech emotion features can be effectively learned utilizing the proposed deep learning approach based on Gammatone auditory filterbank.**

I. INTRODUCTION

Speech emotion recognition (SER) has been a hot topic with a wide range of applications, especially in the field of human computer interaction (HCI). Nevertheless, finding the distinguished feature set for SER is still a challenge due to the different expression types, cultures or context.

Recently, as a most pervasive machine learning method, deep learning has become the best way to find the distinguished feature. There are mainly three kinds of research methods, which are used to recognize the emotion in speech based on deep learning: deep neural networks (DNN) with traditional acoustic features [1, 2], convolutional neural networks (CNN) with spectrogram [3, 4] and CNN and/or recurrent neural networks (RNN) with raw waveform [5, 6, and 7]. Many studies emphasize firstly traditional hand-tuned acoustic features such as prosodic features, voice quality features and spectral features using DNN to find the salient features relevant to emotional speech. Secondly, CNN has been used successfully for a variety of computer vision tasks; hence, quite a few studies recognize the emotion using CNN by converting speech into a special image named spectrogram to build texture and structure representations [8, 9]. Thirdly, end-to-end SER from raw waveform is a new research direction on basis of the strong spatio-temporal feature learning abilities of deep neural networks such as CNN and RNN.

Additionally, computational auditory models [11] based on human auditory system are used in feature extraction for speech emotion recognition tasks [12, 13]. Cochlear filterbank plays a crucial role in computational auditory models, and Gammatone auditory filterbank is a typical cochlear filterbank [14]. In this paper, we put forward to a new method for emotion recognition based on Gammatone auditory filterbank using multichannel parallel convolutional recurrent neural networks (MPCRNN) from raw waveform. We firstly get the Gammatone auditory filterbank from raw audio, and then feed into multichannel parallel CNN after preprocessing in each Gammatone channel. After that, we employ long short-term memory (LSTM), which is a kind of RNN, to get segment-level probability distribution from different channels. Ultimately, we get the utterance level statistic features from segment level features, and then put it into support vector machine (SVM) classifier to predict the emotion state of each utterance.

The rest of the paper is organized as follows. In section 2, some of the main related works are reviewed, and our research is put forward on SER using MPCRNN based on the cochlear filterbank. To do this, MPCRNN architecture of emotion recognition from speech is proposed in section 3. Additionally, experimental settings and results analysis are presented in section 4. At last, we conclude the paper by indicating further prospects of our work in section 5.

II. RELATED WORKS

SER is a challenging task since it is unclear what kinds of features are able to reflect the characteristics of human emotion from speech [2]. Researchers mainly draw special attention to extract salient features from traditional acoustic features for different emotion recognition tasks. Many traditional acoustic features have been investigated, such as prosodic features (duration, F0, energy, zero-crossing rate and speaking-rate features), voice quality features, formant features and spectral features (LPC, MFCC and LPCC features) [15].

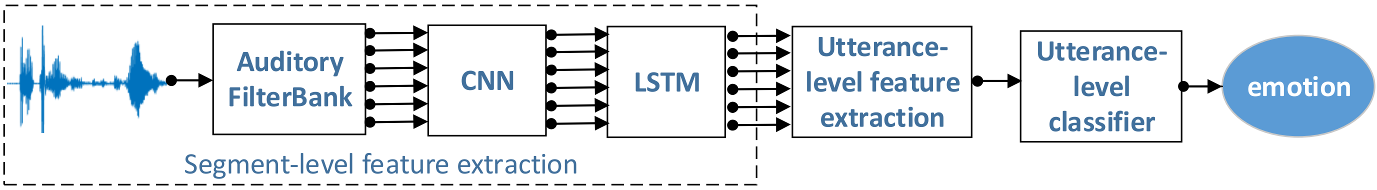

Auditory FilterBank
CNN
LSTM
Utterance-level feature extraction
Utterance-level classifier
emotion
Segment-level feature extraction

Fig. 1: MPCRNN architecture for SER

Additionally, some generative models, such as Gaussian mixture models (GMM) and Hidden Markov models (HMM), are employed to find out the distribution of these acoustic features. Different classification algorithms, such as Bayes, SVM et al., are put forward to recognize emotion by means of extracting effective features from traditional acoustic features.

As deep learning has become the best way to find the distinguished feature, many studies focus on SER using DNN from acoustic features. Han et al. [1] extracted firstly the segment-level emotion state distributions utilizing the traditional features based on DNN model, and employed an extreme learning machine (ELM) to identify utterance-level emotions.

As CNN has been used successfully for a variety of computer vision tasks, several studies have been carried out for the task of SER from spectrogram, which is a kind of visual representation converted from speech. Mao et al. [3] and Huang et al. [4] achieved good performance of speech emotion recognition by trying to learn salient feature maps from spectrogram of speech using an autoencoder followed by CNN.

SER can be treated as a classification problem based on speech sequences. Some studies deal with it as a sequence classification problem employing RNN or LSTM model. Lim et al. [16] and Chernykh et al [17] used deep recurrent neural networks to train on a sequence of acoustic features calculated over small speech intervals. Lee et al. [18] extract high-level representation of emotional states with regard to its temporal dynamics.

Recently, end-to-end speech emotion recognition from raw waveform is a new research trend on basis of the strong spatio-temporal feature learning abilities of deep learning methods. Trigeorgis et al. [5] trained a deep convolutional recurrent neural network to predict dimensional emotions (arousal and valence).

Additionally, Wu et al. [11] and Zhu et al. [12] utilized auditory filterbank for emotional speech analysis. The representation of auditory filterbank captures both acoustic frequency and temporal modulation frequency components, thereby conveying important information for human speech perception. This auditory-inspired band-pass filterbank is constructed as an auditory filterbank by using infinite impulse response filters or finite impulse response filters.

Gammatone auditory filterbank is a group of filters for the cochlea simulation. It is commonly used to simulate the motion of the basilar membrane within the cochlea as a function of time, in which the output of each filter models the frequency response of the basilar membrane. The impulse response of a Gammatone filter is highly similar to the magnitude characteristics of a human auditory filter [14]. Therefore, we attempt to recognize emotion from speech by combining deep learning and human auditory characteristics.

III. MODEL DESIGN

In this section, we specifically emphasize deep learning algorithm based on auditory features to learn discriminative features for SER from raw waveform. We firstly introduce the details of MPCRNN model for SER, and then present the methods for segment-level features and utterance-level emotion.

A. *MPCRNN architecture for SER*

Fig. 1 shows the MPCRNN architecture of the SER system. The first stage is to extract the segment-level robust and compact features from raw audio. The speech signal is firstly segmented into finite length chunks. To mimic the function of basilar membrane, multichannel auditory features are extracted based on Gammatone auditory filterbank in each segment. The multichannel auditory features are subsequently processed to obtain a compact representation of the most salient acoustic characteristics for each channel signal in parallel. Hence, we employ parallel 1-d convolution for each channel in CNN operation, and then feed each channel data into LSTM as a sequential task to get the relations of each channels. We finally get the emotion probability distribution for each segment using MPCRNN model.

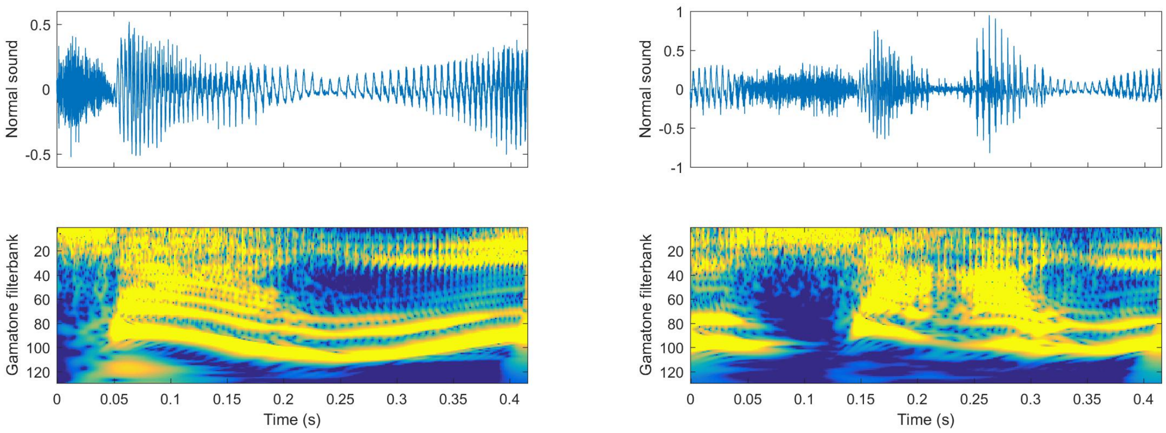
The second stage is to extract the utterance-level statistical features from the different segments belonged to the same utterance, and feed into a SVM classifier to determine the emotional state of the whole utterance.

B. *Segment-level feature extraction*

For extracting the segment-level features, we firstly segment and filter out the raw waveform followed Gammatone auditory filterbank, and train subsequently a MPCRNN to predict the probability distribution of each emotion state.

1) Segmentation and filter for the raw waveform

For segment-level feature extraction, we firstly segment each wav file into 415ms-duration segments. For comparing the traditional method, we get 40 frames for each segment, which includes 25ms windows and 10ms shift.



Happy

Angry

Fig. 2: The graphical representations of Gammatone auditory filterbank for different emotions

The energy of each segment y is the sum of square about each sampling value y_i, as shown in Eq. (1). Moreover, according to the energy of segments, all segments are arranged from lowest to highest.

$$\text{energy}(y) = \sum_{i=1}^{n} {y_i}^2 \tag{1}$$

There are many segments with low energy, so that we cannot perceive any emotion in these segments.

Therefore, we set a threshold value and filter out the segments whose energies are less than the threshold value in accordance with the subjective listening experiments.

2) Gammatone auditory filterbank

Gammatone auditory filterbank models well the basilar membrane motion of human auditory system. The impulse response of a Gammatone filter is the product of a Gamma distribution and a sinusoidal tone. The bandwidth of each filter is described by an equivalent rectangular bandwidth (ERB), which is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea.

$$\text{gt}(t) = \text{A}t^{N-1}\exp(-2\pi b_f \text{ERB}(f_0)t)\cos(2\pi f_0\, t) \tag{2}$$

As shown in Eq. (2), where A, b_f and N are parameters, and $\text{A}t^{N-1}\exp(-2\pi b_f \text{ERB}(f_0)t)$ is the amplitude term represented by the Gamma distribution, f_0 is the center frequency of the filter, and $\text{ERB}(f_0)$ is an equivalent rectangular bandwidth in $f_0(t)$.

Figure 2 shows graphical representations of Gammatone auditory filterbank for different emotions, which are came from the second segment of wangzheangry201.wav and wangzhehappy201.wav with the same sentence. As shown from the upper part of the figure, we find that the speaking rate is faster and the energy is higher with feelings of angry compared to that of happy. Additionally, we find that graphical representations of Gammatone auditory filterbank for different emotions are different with 128 channels filterbank whose center frequencies equals to 600Hz from the lower part of the figure. In other words, different emotions are reacted with different frequency channels in human auditory system.

3) Emotion state probability distribution of each segment using MPCRNN

After adopting the Gammatone auditory filterbank from raw audio, we preprocess each Gammatone channel with zero mean and unit variance, and then feed into MPCRNN. The normalisation ensures that the each channel can catch its own characteristic using the same super parameters in MPCRNN. As shown in Fig. 3, we employ parallel 1-d convolution for each channels in CNN operation, and then feed each channel data into LSTM as a sequential task.

For the CNN part, we use two-layer CNN model to extract different features. The S-Conv layer extracts fine scale spectral information with short window from the high sampling rate signal, while the L-Conv layer extracts more long-term characteristics of the speech with long window. The max pooling operation is employed in each layer.

For the LSTM part, we consider the multichannel convolutional data as a sequence datum and feed the data into two-layer LSTM model. Additionally, fully connected layer is followed by LSTM, which maps the hidden node number (128) into six different emotions: happy, fear, angry, sad, surprise and neutral. The softmax function is then employed to get the probability distribution of each emotion for each segment. At last, the sequence of probability distribution over the emotion states is generated from the segment-level MPCRNN.

Given the sequence of probability distribution over the emotion states generated from the segment-level multichannel parallel convolutional networks, we can form the emotion recognition problem as a sequence classification problem.

C. Utterance-level features

The probability of each segment changes across the whole utterance. Different emotions dominate different regions in the utterance. The true emotion for this utterance is the prominent segment computed from statistics of the segment-level probabilities.

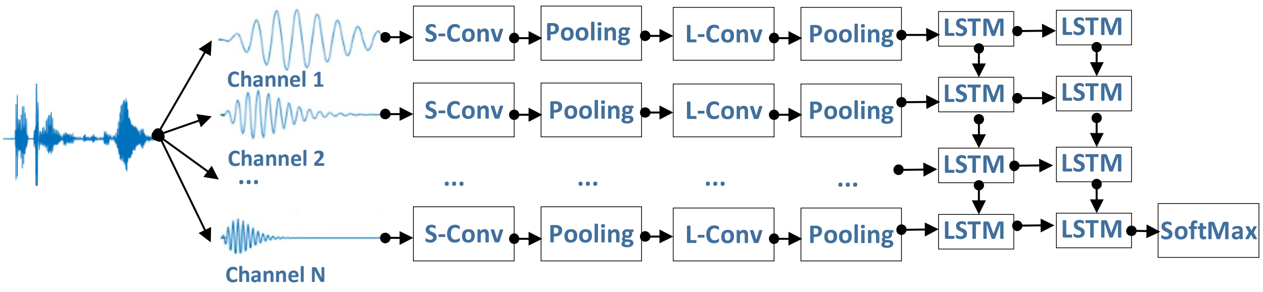


Fig. 3: Segment level features extraction using CRNN based on Gammatone auditory filterbank from raw waveform

In this paper, our experiments are based on the hypothesis that the emotion states of all segments belonged to a certain utterance are the same with the emotion state of this utterance in the training phase. Hence, in the training phase, we assign the same label to all the segments in one utterance. Furthermore, since not all segments in an utterance contain emotional information and it is reasonable to assume that the segments with highest energy contain most prominent emotional information, we only pick out segments with the highest energy in an utterance as the training samples.

The features in the utterance-level classification are computed from statistics of the segment-level probabilities. The maximal, minimal and mean of segment-level probability of the kth emotion over the utterance, respectively. The segment number of each utterance is different with the range from one to eleven.

The MPCRNN has six outputs corresponding to six different emotions. Finally, eighteen utterance-level statistical features are computed with three statistical features for each emotion state and six different emotions totally.

The utterance-level statistical features are fed into a classifier for emotion recognition of the utterance. MPCRNN provides good segment-level results, which can be easily classified with a simple classifier. Therefore, we use a SVM classifier with basic statistical features to determine emotions at the utterance-level. In the testing phase, we get the segment-level probabilities distribution using softmax function, and utterance-level emotions are predicted by means of statistic of the segment-level probabilities.

IV. Experimental Result

A. Experimental data processing

We develop MPCRNN as a fast and optimized algorithm for speech emotion recognition based on Gammatone auditory filterbank. We carried out experiments on CISIA emotional speech database. CISIA is a Mandarin emotional speech database made by Chinese Academy of Sciences. CISIA database comprises totally 9600 recordings from four actors (2 females and 2 males). Recordings for every speaker were made 300 same sentences and 100 different sentences. Each speaker utters 400 sentences with six emotions which are happy, fear, angry, sad, surprise and neutral ("no emotion").

The input signal is sampled at 16 kHz and convert into frames using a 25-ms window sliding at 10-ms each time. So the total length of a segment is 10 ms × 40 + (25 − 10) ms = 415 ms. In fact, emotional information is usually encoded in one or more speech segments whose length varies on factors such as speakers and emotions. According to some studies [24, 25], a speech segment longer than 250 ms has been shown to contain sufficient emotional information.

The threshold value of the energy to filter out the segment is 50. We get 15915 segments as the inputs of MPCRNN from 61938 segments in total. Hence, about 25.7% segments with the highest energy in an utterance are used in the training and the test phase finally.

To get the data from Gammatone auditory filterbank, frequency distributed on ERB scales is between 60 Hz and 6 kHz, and the central frequency f_0 equals to 600 Hz. Meanwhile we apply the four order Gammatone with N equals to four.

B. Hyperparameters for MPCRNN

For training MPCRNN, the S-Conv layer with a 2.5 ms window and 40 kernels in order to extract fine scale spectral information. The L-Conv layer with a 250 ms window and 40 kernels in order to extract more long-term characteristics of the speech. The pool size equals to 2 in the first layer and 10 in the second layer.

We employ parallel convolutional networks with 32 Gammatone channels, and then use two LSTM layers with 128 cells each. For the sequence data with 32 Gammatone channels, we use many-to-one methods to extract the sequence features.

Additionally, for all random weight initializations, we choose L2-regularliser initialization. We employ cross entropy as the objective loss function. We then use Adam gradient descent with the learning rate 1e-5. Moreover, we employ Rectified Linear Unit (ReLU) as the activation function, which brings the non-linearity into networks.

To avoid overfitting in training our networks, we employ dropout [23] as a first measure. Dropout has been specifically proposed for cases where labelled data is scarce. It works by randomly omitting a certain percentage of nodes in the network at training time, while using the full network at test time.

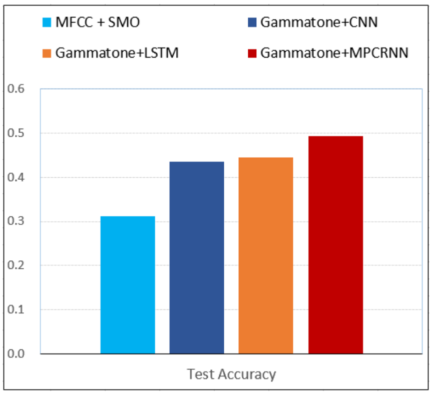


Fig. 4: experiment results on CISIA database

Since deep networks need to be trained on a huge number of training datasets to achieve satisfactory performance, if the original dataset contains limited training data, it is better to do data augmentation to boost the performance. Data augmentation is employed by shifting the original speech audio 300ms and 600ms as a new start point for segmentation as a second measure to avoid overfitting.

C. Results

We train the model in the speaker-independent manner, i.e., we use utterances from three speakers to construct the training datasets, and use the other one speakers for the test. The experiments are performed using Nvidia GTX1080 GPU.

In order to analyze the performance of MPCRNN based on Gammatone auditory filterbank, we also obtain firstly probability distribution for emotional recognition using CNN and LSTM respectively based on Gammatone auditory filterbank. After that, we use a SVM classifier with segment-level statistical features to determine utterance-level emotions.

In addition, we compare our approach with other emotion recognition approach.

MFCC (Mel-frequency cepstral coefficients) is a kind of acoustic features based on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands. OpenSmile as a prevailing emotion recognition toolkit is employed in these approaches to extract the MFCC statistical features [26].

We extract 289 statistics features from 12 MFCC Coefficients for each utterance based on IS09_emotion configure file. We employ these features with SMO (Sequential minimal optimization) classifier, which can get better accuracy than SVM, Naive Bayes and other machine learning methods in CISIA database. In the same way, we achieve in speaker-independent manner.

There are 2400 utterances for each speaker in CISIA database, but some utterances are filtered out as the low energy. Finally, 1520 utterances are remained as the test set[1]. Results obtained for each method are shown in Fig. 4.

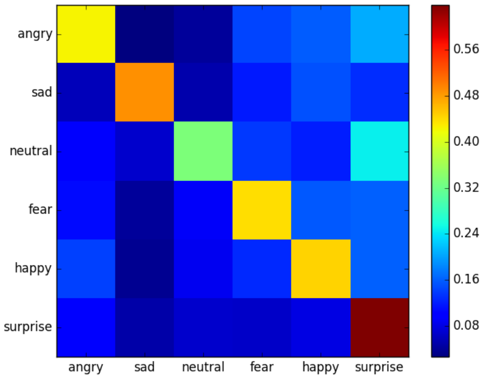


Fig. 5: confusion matrix

In all of the experiments, our study perform better others methods with the accuracy equals to 0.494. We found that MPCRNN outperforms LSTM and CNN by around 10% relatively. The accuracy of MFCC and SMO classifier equals to 0.32 in speaker-independent manner. The proposed approach gives absolute 17.4% better accuracy over the MFCC+SMO approach. Fig. 5 shows the confusion matrix on CASIA. The recognition rate of Surprise is higher than other emotions. A lot of confusion is concentrated between Angry and Surprise. We think this is because there is no distinguishing between Angry and Surprise in valence and arousal space. There is some confusion between Neutral and Surprise. This is because the Neutral has the least samples to extract the salient features.

V. CONCLUSION

In this paper, we studied the recognition of emotional speech by utilizing Gammatone auditory filterbank to train an end-to-end model that combines multichannel parallel convolutional recurrent neural networks. We estimated emotion states for each speech segment in an utterance, constructed an utterance level feature from segment-level estimations, and then employed a SVM classifier to recognize the emotions for the utterance. To our knowledge, this is the first work in literature that applies such a deep learning model based on Gammatone filterbank for speech emotion recognition.

Our experimental results indicate that this approach substantially boosts the performance of emotion recognition from speech signals and it is very promising to use neural networks to learn emotional information based on Gammatone auditory filterbank.

ACKNOWLEDGEMENTS

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026). This work was supported by a Grant-in-Aid for Innovative Areas (No.16H01669) from MEXT, Japan, and by the Secom Science and Technology Foundation. The research was supported partially by the National Basic Research Program of China (No. 2013CB329301), and the National Natural Science

[1] Class distribution: angry: 346; fear: 237; happy: 218; neutral: 117; sad: 237; surprise: 365

Foundation of China (No. 61233009). The study was supported partially by JSPS KAKENHI Grant (16K00297).

REFERENCES

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *in Proceedings of interspeech 2014*, pp. 223-227.

[2] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," *in Proceedings of ACII 2015*.IEEE, 2015,pp 827-831

[3] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using- convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014.

[4] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," *in Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 801-804.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, e E. Marchi, M. A. Nicolaou, B. Schuller,and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *in Proceedings of IEEE ICASSP 2016*. IEEE, 2016, pp. 5200-5204

[6] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," *in Proceedings of ICML2014. Vol. 14, pp. 1764-1772.*

[7] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv*:1408.5882, 2014.

[8] J. H. Mcdermott, E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," Neuron 71.5(2011):926-940.

[9] J. H. Mcdermott, M. Schemitsch, E. P. Simoncelli, "Summary statistics in auditory perception," Nature Neuroscience 16.4(2013):493-8.

[10] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech communication, 2011, vol. 53, no. 5, pp. 768-785.

[11] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation Spectral Features for Predicting Vocal Emotion Recognition by Simulated Cochlear Implants", in Proceedings of interspeech 2016, pp. 262-266.

[12] Z. Zhu, R. Miyauchi, Y. Araki, M.Unoki, "Recognition of vocal emotion in noise-vocoded speech by normal hearing and cochlear implant listeners," Journal of the Acoustical Society of America 140(2016).

[13] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," Speech Communication vol. 27, no. 3, pp. 261-279, 1999.

[14] B. Zhang, C. Quan, and F. Ren, "Study on CNN in the recognition of emotion in audio and images," Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, 2016, pp. 1-5.

[15] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition 44.3(2011):572-587.

[16] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in Proceedings of APSIPA 2016, pp. 1-4.

[17] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion Recognition From Speech With Recurrent Neural Networks," arXiv preprint arXiv:1701.08071, 2017.

[18] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," in Proceedings of interspeech 2015, pp. 1537-1540.

[19] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in Proceedings of IEEE ICASSP 2015. IEEE, 2015: pp. 4624-4628.

[20] A. Sangari and W. Sethares, "Convergence Analysis of Two Loss Functions in Soft-Max Regression," in IEEE Transactions on Signal Processing, vol. 64, no. 5, pp. 1280-1288, 2016.

[21] G. Yu, E. Postma, H. X. Lin, and J. van den Herik, "Speech Emotion Recognition with Log-Gabor Filters," Proceedings of the 8th International Conference on Agents and Artificial Intelligence 2016, pp. 446-452.

[22] B. Milde and C. Biemann, "Using representation learning and out-of-domain data for a paralinguistic speech task," in Proceedings of interspeech 2015, pp. 904-908.

[23] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing coadaptation of feature detectors," arXiv preprint, pp. 1-18, 2012

[24] Y. Kim and E. Mower Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in Proceedings of IEEE ICASSP 2013. IEEE, 2013, pp. 3677-3681.

[25] E. Mower Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in Proceedings of IEEE ICASSP 2013. IEEE, 2013, pp. 3682-3686.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," ACM International Conference on Multimedia ACM, 2010, pp. 1459-1462.