Auditory Physiology Based Emotion Recognition System

Wei-Ting Lin, Kai-Wen Liang, and Pao-Chi Chang
Department of Communication Engineering
National Central University
Taoyuan City, Taiwan
{waitinglincomm@gmail.com, kwistron@gmail.com, pcchang@ce.ncu.edu.tw}

***Abstract*—This research proposes an emotion recognition system that lies on mimicking the human auditory perception model based on two-stage auditory module, including the early stage, and the cortical stage. It stimulates the output auditory spectrogram of the early stage by MFCCs feature extraction, employs Gabor filter banks as spectral-temporal analysis in the cortical stage, and embeds CNN structure in the final stage to refine and fusion pattern of features. The experimental results show that the proposed method outperforms the existing methods in speech emotion recognition.**

Keywords—Neural system, CNN, Emotion Recognition

I. INTRODUCTION

Researches about speech and emotion recognition have been undergone for years. In spectral processing, Mel-Frequency Cepstral Coefficients (MFCCs), and Perceptual Linear Prediction features (PLPs) are classic methods for sound feature extraction. While, from a physiological point of view, researches show neurons in the primary auditory cortex of mammals can extract spectro-temporal patterns from the inputting sound signal. Different sorts of spectro-temporal characteristics for automatic speech recognition (ASR) have thus been developed. Those approaches to acquire spectro-temporal features from speech, including derive spectro-temporal receptive fields (STRFs) from neural network, employ patches with 2D discrete cosine transform, and adopt Gammatone filterbank or Gabor features to improve speech recognition and robust features acquiring for analysis. Recently, as the prevalent of machine learning, deep learning networks also provide numerous ways for distinguishing features in the spectrum.

In our study, we would like to create an emotion recognition system based on human's two-stage auditory model, the early stage and the cortical stage, developed by the Neural Systems Laboratory (NSL) [1]. The mechanism in the Early Stage, including cochlear, hair cell, and lateral inhibitory network (LIN) is described through a succession of formulations as constant-Q bandpass filters, a non-linear compression module, and an envelope extractor. The cortical stage acts as the phase of spectro-temporal analysis. The spectro-temporal receptive fields (STRF) of these filters are proposed to capture speech characteristic in different Ifrequencies and angles.

We employ MFCCs, Gabor filter banks and Convolutional Neural Network(CNN) to stimulate the human perception of the Early Stage, the Cortical Stage, and the unknown message transmission mechanism among neurons in brain

II. GABOR FILTER BANK GEATURES

Frequency and orientation representations obtained by Gabor filters shown applicable on enhancing detection and recognition in different types of phenomena. A general type of 2D Gabor filters can be defined by a sinusoidal wave multiplied by a Gaussian function as eq. 1.

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \quad (1)$$

$$x' = x\cos\theta + y\sin\theta,\ y' = -x\sin\theta + y\cos\theta$$

In our research, we implement a type of Gabor filter which satisfies the neurophysiological constraints for simple cells:

$$\psi(x; \omega, \theta, K) = \left[\frac{\omega^2}{4\pi K^2}\exp\{-(\omega^2/8K^2)[4(x\cdot(cos\theta, sin\theta))^2 + (x\cdot(-sin\theta, cos\theta))^2]\}\right]$$

$$\times\ [\exp\{iwx\cdot(cos\theta, sin\theta)\}exp(K^2/2)] \quad (2)$$

We take the study[2] as a reference to design a set of suitable parameters which results the enhancement of the robustness, and create 24 2D Gabor filter banks with parameters of $\theta = \{\frac{7\pi}{8}, \frac{6\pi}{8}, \frac{4\pi}{8}, \frac{2\pi}{8}, \frac{\pi}{8}, 0\}$(rad), $\omega = \{0.188, 0.37, 0.75, 1.57\}$(rad/s), as Fig. 1.

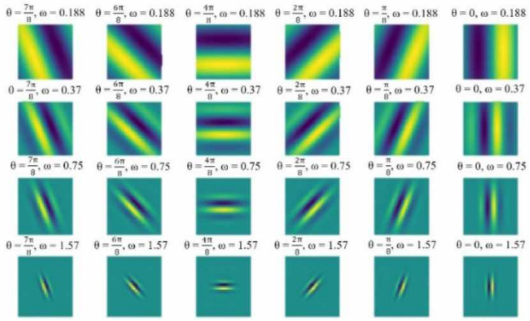

Fig. 1. Real Components of Gabor Filter Banks

III. PROPOSED SYSTEM MODEL

Our emotion recognition system model includes three parts: The Early Stage, The Cortical Stage and CNN structure, as Fig.2

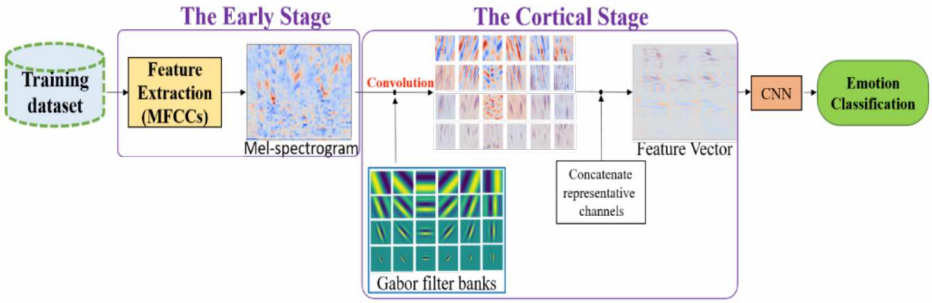

Fig. 2. Emotion recognition system model

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

A. *The Early Stage*

As MFCCs feature extraction process also takes the mechanism of human auditory system into account and is highly correspond to the module in the Early Stage of the auditory physiology proposed by NSL[1], we adopt MFCC to stimulate the early stage, shown as in Fig.3.

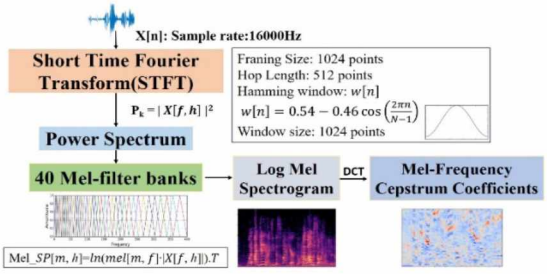

Fig.3. MFCCs feature extraction

B. *The Cortical Stage*

After generating the auditory spectrogram, we implement log Mel-spectrogram with Gabor filter bank, as in Fig.1, to mimic the cortical stage. We can then obtain 24 spectro-temporal spectrograms, each extract different frequency and angle characteristics from spectrograms, shown as in Fig. 4.

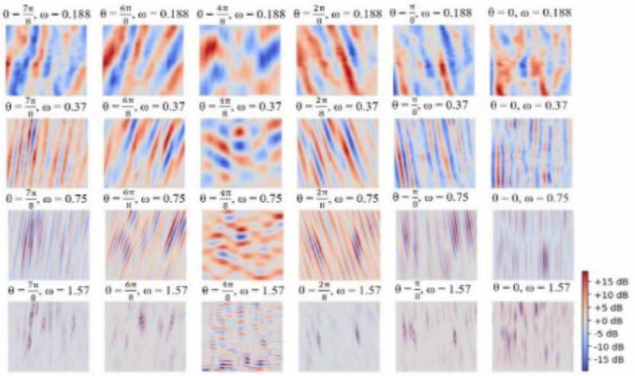

Fig. 4. Spectro-temporal spectrograms

To streamline the features within 24 spectro-temporal spectrograms, we select the representative channels from each: 9, 10, 11 frames as features of low frequency, 19, 20, 21 frames as features of center frequency, and 29, 30, 31 frames as features of high frequency. We concatenate the representative channels of each spectro-temporal spectrogram and get the resulting 216-dimensional output represented as the Feature Vector.

C. *Convolutional Neural Network*

Thanks to the self-optimized traits of machine learning model that is suitable to be applied to unknown message-delivering mechanisms of neurons, we implement Convolutional Neural Networks (CNN) on the Feature Vector, shown as in Fig. 5, in the final stage.

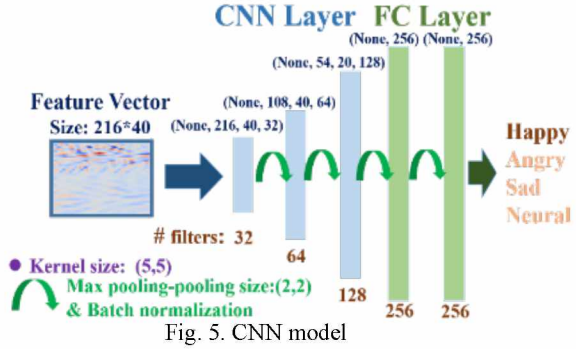

Fig. 5. CNN model

IV. Experiments And Eiscussion

A. *Speech Data*

We adopt the database of Interactive Emotional Dyadic Motion Capture (IEMOCAP) [3]. We consider 4861 speech data in total consists of happy, anger, sadness and neutral, four categories of emotion.

B. *Experimental results and discussion*

TABLE 1. shows the comparison of accuracy between our work and previous researches adopt IEMOCAP as database.

TABLE 1.

Model	Accuracy
Lakomkin [4]	56%
Neumann [5]	56.10%
Tripathi and Beigi [6]	62.72%
Lee and Tashev [7]	62.85%
Ours	**63.1%**

When generating Feature Vector, we need to figure out which merging order of representative channels stands for the best spectrum feature. We found the best combining order would be- First, concatenate representative channels chosen from each spectro-temporal spectrogram. Then combine them in an order of angle and frequency from small to large, respectively. Moreover, in the final stage, rather than adopting complex or combined deep learning model, we found such a simple structure is capable enough to extract essential features from the inputting pattern.

V. Conclusion

In this research, to develop a model more proximate to human auditory physiology, a two-stage methodology is adopted. Our method can achieve the accuracy up to 63.1% for IEMOCAP dataset, which is superior to previous studies. This work contributes to show the Gabor filter banks are helpful in extracting different frequency and angle characteristics among the auditory spectrum, and the combination of the CNN structure and two-stage auditory perception model performs well on the recognition. To sum up, we consider the human perception is worthwhile taking into account when working on emotion estimation researches.

VI. Reference

[1] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," The Journal of the Acoustical Society of America, vol. 118, no. 2, pp. 887–906, 2005.

[2] H. Lei, B. Meyer, N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition," in Proc. ICASSP, 2012.

[3] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[4] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing neural speech representations for auditory emotion recognition," arXiv preprint arXiv:1803.11508, 2018.

[5] S. Tripathi and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," arXiv preprint arXiv:1804.05788, 2018.

[6] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," arXiv preprint arXiv:1706.00612, 2017.

[7] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 2015

978-1-7281-9802-6/20/$31.00 ©2020 IEEE