Speech Emotion Recognition Based on Joint Self-Assessment Manikins and Emotion Labels

Jing-Ming Chen
Dept. of Communication Engineering
National Central University
Taoyuan City, Taiwan, ROC
jmchen.vaplab@gmail.com

Pao-Chi Chang
Dept. of Communication Engineering
National Central University
Taoyuan City, Taiwan, ROC
pcchang@ce.ncu.edu.tw

Kai-Wen Liang
Dept. of Communication Engineering
National Central University
Taoyuan City, Taiwan, ROC
kwistron@gamil.com

***Abstract*—In this work, we propose a system for speech emotion recognition based on regression models and classification models jointly. This speech emotion recognition technology can achieve the accuracy of 64.70% in the dataset of script and improvised mixed scenes. The accuracy can be up to 66.34% in the dataset with only improvised scenes. Compared to the state-of-art technology without the mental states, the accuracy of the proposed method is increased by 2.95% and 2.09% respect to improvised and mixed scenes. The results show that the characteristics of mental states can effectively improve the performance of speech emotion recognition.**

Keywords—Speech emotion recognition, Self-Assessment Manikin, Deep learning, Convolutional recurrent neural network

I. INTRODUCTION

With the advancement of technology, the impact of artificial intelligence (AI) to modern lives has been increased gradually. Applications such as smart home appliances, chatbots, etc., are not only getting popular, but also bring a lot of convenience to our daily lives. Most of these applications are benefit from the practical interaction between human and machines. Speech emotion recognition is one of the good examples that AI technology makes robots understand the emotions of human beings to react relatively and has attracted a lot of attention in the research field lately.

The mainstream technology behinds most of these products is deep learning technology. Deep learning technique has evolved the research field of image processing. Techniques based on deep learning for issues such as the pedestrian tracking and the face recognition have been continuously developed with very good performance. On the other hand, in the research field of acoustic processing, deep learning techniques were also applied to develop systems to sound event detection, sound scene classification, speaker recognition, etc., all of which can be realized with better performance than the traditional methods.

II. EMOTIONAL RELATED BACKGROUND

For human emotions, in addition to the commonly used characterization categories, such as happiness, sadness, sadness, etc., mental state can also be used to judge emotions, such as positive degrees or excitement degree are help to distinguish emotions. We hope using deep learning techniques to achieve emotion recognition by combining these psychological characteristics with emotion labels.

Human's emotions can be categorized as happiness, sadness, angry, etc., in a common sense. Mentally, we can further make a judgement to a kind of emotion with the help of the status of the current emotion. That means degrees of positive or excitement can also be considered and supposed to be beneficial to emotion recognition.

Speech emotion recognition is the technique which enables a computer to identify the emotion by analyzing the acoustic characteristics of the speech. Generally, short-time Fourier transform is used to extract the acoustic features, that is, the spectrum of audio or speech data. And to many of the state-of-art researches , deep learning techniques are used to design the neural networks that can achieve the task of emotion recognition by classifying the input data. There are two orientations of studies to emotion recognition, discrete emotion theory and emotional dimension model. In general, the discrete emotion theory is commonly used due to the fact that it can clearly classify emotions into different categories, such as happy, sad, angry, scared, surprised, disgusted, etc. On the other hand, the emotional dimension model is based on different essences of emotions represented by differences in degree, the general model includes Valence and Arousal dimensions, as well as more dimensions of the model like the Plutchik model [1], PAD emotional state model [2] and so on.

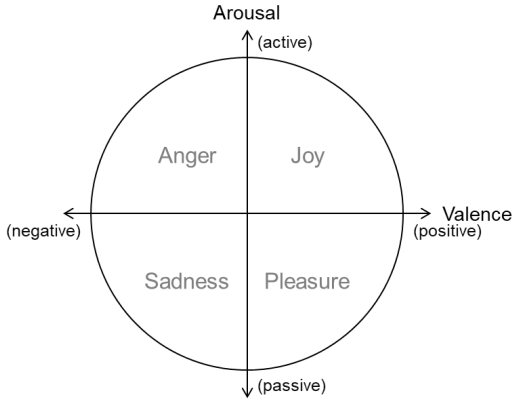

Fig. 1. Two dimension emotion model of Valence and Arousal.

The PAD emotional state model is a three-dimensional model proposed by M. Albert to measure the emotional state of the mind, which include Pleasure, Arousal and Dominance. Pleasure represents the positive and negative of emotions, Arousal respond to the excitation level is high or low, and Dominance represents the speaker's emotional control performance is strong or weak.

Self-Assessment Manikin (SAM) [3] is a method for marking the degree of emotional state, proposed by M. M. Bradley and P. J. Lang in 1994, and its scale is shown in Figure 2. The evaluation category of this method includes three dimensions of the PAD emotional state model: Valence, Arousal and Dominance, each emotional state is divided into five levels according to the degree of difference.

978-1-7281-5606-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ISM46123.2019.00073

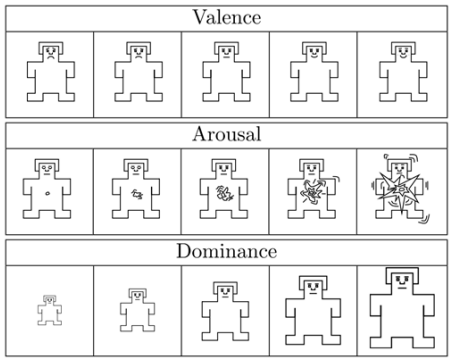

Fig. 2. Self-Assessment Manikin assess for the valence, arousal, and dominance in 5-point rating scale range.

III. Proposed Method

Figure 3 shows the proposed system for speech emotion recognition. This system consists of three parts: feature preprocessing, Self-Assessment Manikin regression model and speech emotion classification model.

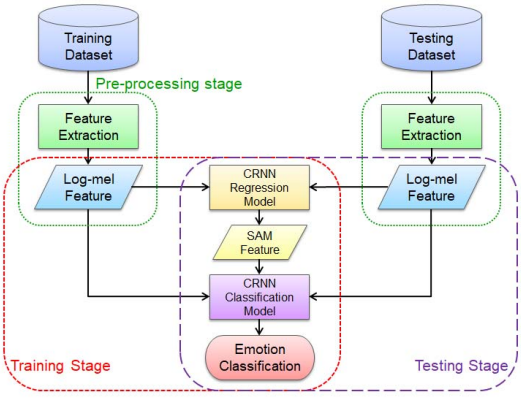

Fig. 3. System architecture of speech emotion recognition.

A. Feature Preprocessing

In feature extraction, we used the log-Mel spectrogram, Delta spectrogram, and Delta-Delta spectrogram as inputs to the neural network.

Log-Mel spectrogram extraction process is shown in Figure 4. First, the signal samples are segmented into frames of 25 ms with 10 ms overlap. Then a Hamming window is multiplied and the 512-bin Fast Fourier Transform (FFT) is computed for each frame. Finally, the spectrogram is applied to 40-band Mel-scale filter-bank and calculated by logarithm power.

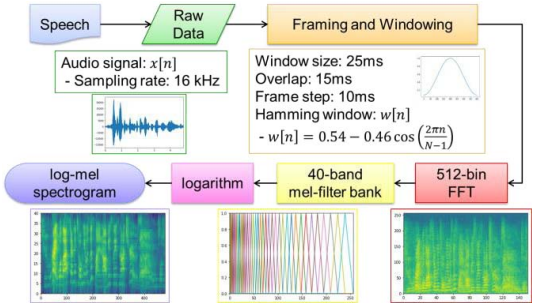

Fig. 4. Flow chart of feature extraction in log-Mel spectrogram.

Like [4], we also use Delta and Delta-Delta features as input, they are calculated by the following formula (1) and (2). surveyThus, the dimensions of our input has 3 channels (log-Mels, Delta, and Delta-Delta), and the size of each channel is N (frames) × 40 (features).

$$m_i{}^d = \frac{\sum_{t=1}^{T} t(m_{i+t} - m_{i-t})}{2\sum_{t=1}^{N} t^2} \quad (1)$$

$$m_i{}^{dd} = \frac{\sum_{t=1}^{T} t(m_{i+t}{}^d - m_{i-t}{}^d)}{2\sum_{t=1}^{T} t^2} \quad (2)$$

B. Convolutional recurrent neural network

This study used convolutional recurrent neural network and was divided into two parts: the Self-Assessment Manikin regression model and the speech emotion classification model (Figure 5).

The Self-Assessment Manikin regression model contains shared layers and task-specific layers. The shared layers consist of multi-layer convolutional layer, bidirectional gated recurrent unit layer, and a fully connected layer. The task-specific layers are composed of multi-layer fully connected layers to predict the values of valence, activation, and dominance.

The speech emotion classification model also includes a multi-layer convolutional layer, bidirectional gated recurrent unit layer, and a multi-layer fully connected layer. The first fully connected layer is concatenate with the first layer of the fully connected layer in the regression model to learn the characteristics of the mental emotional state, and finally output the probability of each emotional category.

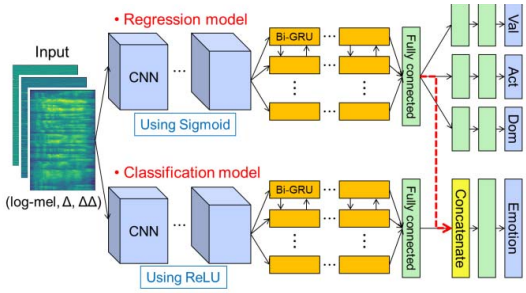

Fig. 5. The architecture of proposed speech emotion recognition model.

IV. Experiments

A. Experimental Setup

We using Interactive Emotional Dyadic Motion Capture (IEMOCAP) [5] dataset for our proposed system, it contains about 12 hours of audio and video data, and has ten actors performed five sessions, which were completed by a male actor and a female actor in each session. In addition to script scenario, it also has improvise scenario to simulate the emotional situation of reality. In the categories of emotions, there are nine emotions such as anger, sadness, happiness, frustration, fear, surprise, depression, excited and neutral.

In order to consistent with previous work [6], we evaluate the system for four emotions: Anger, Happiness, Neutral, and Sadness. Also, we merge Excitement into Happiness class. Our training data used session 1~4, and test data used session 5.

B. Metrics

The evaluation metrics of the regression analysis can judge the performance of the model based on the correlation between the prediction result and the correct answer. In this study, we use Concordance Correlation Coefficient (CCC) [7] as the indicator of judgment, which is calculated by the formula (3).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2 + \left(\mu_x - \mu_y\right)^2} \quad (3)$$

The value calculated by the consistency correlation coefficient will between -1 and 1, where 1 means the two sequences are completely related, -1 means the two sequences are completely inversely related, and if 0, the two sequences are completely unrelated. For regression analysis, the goal is to make the predicted value close to or completely the same as the expected value, therefore, the effect of the regression analysis is better when the value of the consistency correlation coefficient approaches to 1.

C. Experimental Results

In the architecture selection of the Self-Assessment Manikin regression model, since the recurrent neural network is very sensitive to the weight adjustment during the training process, if the parameters are not properly selected, it is easy to cause the neural network to overfitting. In the preliminary experiment, we tried to train the model using different recurrent neural network layers to find the most suitable recurrent neural network layer. The experimental results are shown in Table I.

TABLE I.

	Valence CCC	Activation CCC	Dominance CCC	Average CCC
1 CNN 1 RNN	0.39498	0.65124	0.48319	**0.50980**
1 CNN 2 RNN	0.39498	0.65727	0.49016	**0.51455**
1 CNN 3 RNN	0.40338	0.66474	0.47234	**0.51349**
1 CNN 4 RNN	0.40150	0.68284	0.50358	**0.52931**

From this experimental result, the more number of layers used in the recurrent neural network, the higher consistency correlation coefficient. In the correlation performance of the three mental state features, the valence is the lowest of the three, the second is dominance, and the highest is activation. As the number of layers increases, there is almost no influence on the degree of valence, but the degree of dominance and activation have an increasing trend.

Next, we select the number of convolutional layers to analyze which layers have a better performance. The experimental results are shown in Table II.

TABLE II.

	Valence CCC	Activation CCC	Dominance CCC	Average CCC
1 CNN 4 RNN	0.40150	0.68284	0.50358	**0.52931**
2 CNN 4 RNN	0.44693	0.67500	0.51268	**0.54487**
3 CNN 4 RNN	0.46440	0.69524	0.51932	**0.55965**
4 CNN 4 RNN	0.46623	0.68170	0.52591	**0.55795**
5 CNN 4 RNN	0.46414	0.68366	0.52635	**0.55805**

The better consistency correlation coefficient obtained from the experimental results when the convolutional layer is three layers. When the convolutional layer is increased from one layer to three layers, the correlation is rising, but increasing the convolution layer upwards will not bring better results, the reason is that when the pooling layer is operated after the convolutional layer, the feature map is downsampled, so although the deep convolutional neural network can learn a wider range of features, it also causes information loss in the frequency domain. Therefore, the Self-Assessment Manikin regression model is suitable for using a three-layer convolution layer.

In the architecture selection of the speech emotion classification model, the different number of convolution layers is also trained to find the best number of convolution layers. The experimental results are shown in Table III.

TABLE III.

	Anger	Happy	Neutral	Sadness	Accuracy
1 CNN 1 RNN	0.75	0.27	0.72	0.64	**58.34%**
2 CNN 1 RNN	0.68	0.36	0.69	0.74	**60.83%**
3 CNN 1 RNN	0.69	0.33	0.73	0.74	**61.75%**
4 CNN 1 RNN	0.59	0.33	0.79	0.71	**61.65%**
5 CNN 1 RNN	0.64	0.40	0.66	0.76	**60.46%**

The experimental results in Table III are consistent with the Self-Assessment Manikin regression model, the results of the three-layer convolution layer is the best. When the number of layers is increased from one layer to three layers, the overall accuracy increases, and the more layers are no more contribution to the accruacy. Therefore, it is most appropriate to select a three-layer convolutional neural network in the log-Mel spectrogram.

Next, we conducted a recursive neural network layer number selection experiment in the speech emotion classification model to observe whether the temporal characteristics are helpful for classification. The experimental results are shown in Table IV.

TABLE IV.

	Anger	Happy	Neutral	Sadness	Accuracy
3 CNN 1 RNN	0.69	0.33	0.73	0.74	**61.75%**
3 CNN 2 RNN	0.63	0.31	0.70	0.79	**60.18%**
3 CNN 3 RNN	0.58	0.32	0.70	0.80	**59.90%**

In this experimental results, increment of the RNN layers is not useful to improve the accuracy but leads to worse result instead. Therefore, temporal characteristics are not very helpful to emotion classification models, so the mental emotional states work differently from the emotional labels.

In this study, the mental-emotional states and emotion labels were trained and combined by two neural networks to achieve better emotional recognition. The experimental results are shown in Table V.

TABLE V.

Model		Accuracy
All scenes	**Baseline [6]**	**55.65%**
	Ours (CRNN)	**61.75%**
	Ours (SAM + CRNN)	**64.70%**
Improvise only	**Baseline [6]**	**62.72%**
	Ours (CRNN)	**64.25%**
	Ours (SAM + CRNN)	**66.34%**

After training the emotion classification model with the Self-Assessment Manikin regression model, the accuracy obtained in the mixed scenario of scripts and improvisations performance is 64.70%, and the accuracy obtained only in the scenario of improvisation performance is 66.34%. Compared with the model without the combination, the combined model has a higher accuracy of 2.95% and 2.09% in mixed and improvisation only scenarios, respectively. The results show that the mental-emotional states are beneficial to emotional classification.

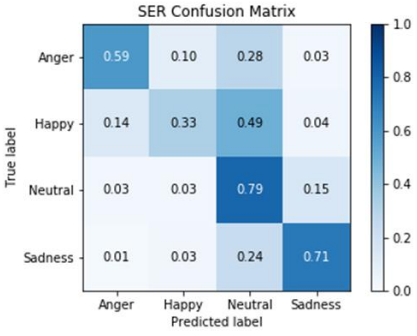

Fig. 6. Confusion matrix of speech emotion classification model in the mixed scenario.

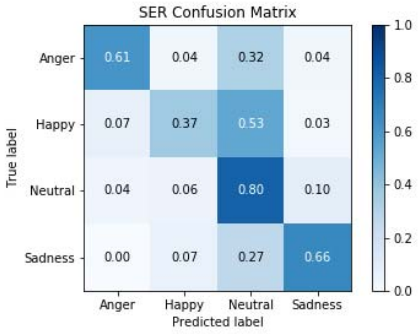

Fig. 7. Confusion matrix of speech emotion classification model with mental emotional states in the mixed scenario.

V. Conclusion

This study proposes a speech emotion recognition technique that combines mental-emotional states and emotion labels trained on convolutional recurrent neural network. In the selection of the neural networks model, the mental-emotional states are predicted by the regression model, and the emotion labels are classified by the classification model.

This speech emotion recognition technique can achieve a good accuracy up to 66.34% in the improvisation scenario. In the scenario of which scripts and improvisations are mixed, the accuracy can also reach 64.70% which is better than the benchmark model.

References

[1] R. Plutchik, "Emotions and life: Perspectives from psychology, biology, and evolution," American Psychological Association, 2003.

[2] P. Valdez, and A. Mehrabian, "Effects of color on emotions," Journal of experimental psychology: General, vol. 123, no. 4, 1994.

[3] M. M. Bradley, and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," Journal of behavior therapy and experimental psychiatry, vol. 25, no. 1, pp. 49-59, 1994.

[4] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, 2018.

[5] 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.

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

[7] I. Lawrence, and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," Biometrics, pp. 255-268, 1989.