SPEECH EMOTION RECOGNITION WITH CO-ATTENTION BASED MULTI-LEVEL ACOUSTIC INFORMATION

Heqing Zou[1], *Yuke Si*[2], *Chen Chen*[1], *Deepu Rajan*[1], *Eng Siong Chng*[1]

[1]Nanyang Technological University, Singapore [2]Tianjin University, China

ABSTRACT

Speech Emotion Recognition (SER) aims to help the machine to understand human's subjective emotion from only audio information. However, extracting and utilizing comprehensive in-depth audio information is still a challenging task. In this paper, we propose an end-to-end speech emotion recognition system using multi-level acoustic information with a newly designed co-attention module. We firstly extract multi-level acoustic information, including MFCC, spectrogram, and the embedded high-level acoustic information with CNN, BiLSTM and wav2vec2, respectively. Then these extracted features are treated as multimodal inputs and fused by the proposed co-attention mechanism. Experiments are carried on the IEMOCAP dataset, and our model achieves competitive performance with two different speaker-independent cross-validation strategies. Our code is available on GitHub.

Index Terms**—** Speech emotion recognition, Multimodal fusion, Multi-level acoustic information, Co-attention mechanism

1. INTRODUCTION

Automatic recognition of emotions finds several applications such as human-computer interaction [1] and surveillance [2]. Some researchers propose to combine acoustic information with textual information and learn high-level context information to help make the final emotion prediction [3]. However, the corresponding transcriptions are not always available for most emotion recognition applications. Besides, the generated text with a current automatic speech recognition (ASR) system could also introduce word recognition errors and interfere with the emotion recognition task. Emotion perception from only audio signals is much easier to implement compared with multimodal emotion recognition with additional textual and visual signals because single audio data is easier to be obtained. Transforming the speech emotion recognition (SER) problem into a multi-level fusion problem by integrating multiple acoustic information is a potentially effective method to utilize the complete audio information.

The vast majority of SER problems involve extracting key audio features like Mel-frequency Cepstral Coefficient (MFCC), Constant-Q Transform (CQT) or constructing the corresponding spectrogram image to treat the problem as an image classification problem [4]. Both MFCC and spectrogram reflect more information of a speech signal in the frequency domain. MFCC can be regarded as a low-level feature based on human knowledge. Spectrogram can be further processed to obtain high-level information through a deep neural network. These methods are intuitive and simple but usually ignore time-domain information of the speech signal.

Various encoders with different architecture details are designed for different acoustic signals, e.g., CNN for spectrogram and CNN/LSTM for MFCC. The acoustic information is mined using a series of CNNs with different kernel sizes in [3]. Some methods propose to introduce a combination of networks to extract acoustic information, e.g., [5] combine LSTM and Gated Multi-features Unit (GMU) to extract both static and dynamic speech signals. In [6], Gao et al propose a domain-adversarial auto-encoder to extract discriminative representations with pre-trained spectrogram information. Extracting features from different sources requires the corresponding source-specific neural networks.

Different types of attention mechanisms have been proposed for processing the extracted features, like the commonly used self-attention [5, 7] and cross-modal attention [8]. For the models with more complex input combinations, new attention mechanisms are introduced. [9] fuses two modalities and then combines the result with another modality using the proposed attentive modality-hop mechanism. In [10], a hierarchical attention-based temporal convolutional network is designed to fuse the inter-channel and intra-channel features for spectrogram images.

In this paper, we introduce three different encoders for multiple levels of acoustic information: CNN for spectrogram, BiLSTM for MFCC and the transformer-based acoustic extracting network wav2vec2 [11] for raw audio signals. With the designed co-attention module, we optimize to get the final wav2vec2 embedding (W2E) after weighting each frame by utilizing the effective information extracted from MFCC and spectrogram features. We concatenate all three extracted features and make the final emotion prediction with this finally fused information. The proposed model surpasses current competitive models on the widely used IEMOCAP dataset with the leave-one-speaker-out and leave-one-session-out cross-validation strategy.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

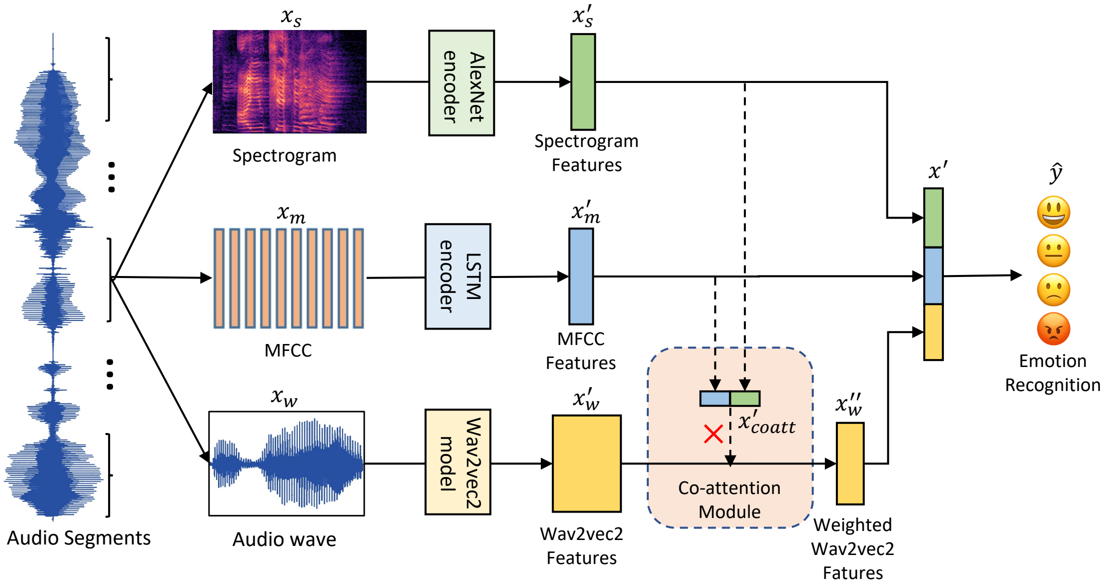


Fig. 1. The overall architecture of our proposed method.

2. PROPOSED METHOD

In this section, we describe our co-attention-based SER system by integrating multiple acoustic information. Fig.1 shows the overall structure of our proposed method. As illustrated, after splitting the raw audio utterance into several segments, three levels of acoustic information (MFCC, spectrogram and W2E) of a segment are introduced to the respective feature encoder networks and fused with the proposed co-attention method for the final emotion recognition.

2.1. Model Overview

We denote the MFCC, spectrogram and wav2vec2, which are obtained from the same audio segment, as $x_m \in \mathbf{R}^{T_m \times D_m}$, $x_s \in \mathbf{R}^{T_s \times D_s}$ and $x_w \in \mathbf{R}^{T_w \times 1}$, respectively. The extracted MFCC features x'_m and spectrogram features x'_s are concatenated and transformed with linear layers to get the weights for different frames of wav2vec outputs x'_w. After multiplication with these generated weights, we get the final W2E vector from the raw wav2vec outputs. The final obtained W2E x''_w are concatenated with the previous MFCC features x'_m and spectrogram features x'_s for the final emotion recognition task. The generated weights of wav2vec frames from MFCC and spectrogram features and the final feature combination are denoted as x'_{coatt} and x', respectively. The target of the data is denoted by y and the final prediction is denoted as $\hat{y}$.

2.2. Learning with Multi-level Acoustic Information

Here we define the multi-level acoustic information as the combination of the human knowledge based low-level MFCC, deep learning based high-level spectrogram and W2E, thus to cover characteristics of the speech signal in both frequency and time domain. MFCC sequence is processed by a bidirectional LSTM with a dropout of 0.5 and flattened. The flattened vector is input to a linear layer with ReLU as an activation function with a dropout of 0.1 to obtain

$$x'_m = f_m(BiLTSM(x_m)) \tag{1}$$

where $x'_m \in \mathbf{R}^{D'_m}$.

The spectrogram image is first reshaped for the pre-trained AlexNet. A similar operation as for MFCC features is conducted on the AlexNet extracted features to obtain

$$x'_s = f_s(AlexNet(x_s)) \tag{2}$$

where $x'_s \in \mathbf{R}^{D'_s}$.

Raw audio segments are sent directly to the corresponding wav2vec2 processor and wav2vec2 model to get the target raw wav2vec2 outputs as

$$x'_w = Wav2Vec2(x_w) \tag{3}$$

where $x'_w \in \mathbf{R}^{T'_w \times D'_w}$.

2.3. Co-attention-based Fusion

Considering that all three acoustic information sources play a similar role in the final emotion prediction, we use the correlation among them to guide the feature adaptation. Generally, the last frame or the average of the wav2vec2 output is used to represent the wav2vec2 features. It is obvious that we lose some effective information among the sequence dimension.

Table 1. Performance comparison with 5-fold leave-one-session-out [12,7,13] and 10-fold leave-one-speaker-out [14,15,16,17,18,5] cross-validation strategy on IEMOCAP.

Model	WA	UA
CNN-ELM+STC attention[12]	61.32	60.43
Audio$_{25}$[7]	60.64±1.96	61.32±2.26
IS09 - classification [13]	68.1	63.8
Ours	**69.80**	**71.05**
RNN(prop.)-ELM[14]	62.85	63.89
3D ACRNN[15]	-	64.74±5.44
BLSTM-CTC-CA[16]	69.0	67.0
CNN GRU-SeqCap[17]	72.73	59.71
CNN_TF_Att.pooling[18]	71.75	68.06
HNSD[5]	70.5	72.5
Ours	71.64	**72.70**

Here, we introduce a kind of co-attention module to combine different frames of W2E with frame weights generated by the features of MFCC and spectrogram features.

Firstly, we create a 1-dimension matrix from MFCC features x'_m and spectrogram features x'_s with a transformation layer given by

$$x'_{att} = f_{att}(x'_m \oplus x'_s) \tag{4}$$

where $x'_{att} \in \mathbf{R}^{1\times T'_w}$.

The wav2vec2 outputs are multiplied with the previous generated weights to get the final weighted wav2vec2 features as

$$x''_w = (x'_{att} \cdot x'_w)^T \tag{5}$$

where $x''_w \in \mathbf{R}^{D'_w}$.

The final MFCC, spectrogram features and the weighted W2Es are concatenated and the speech emotion prediction is written as

$$\hat{y} = f(x'_m \oplus x'_s \oplus x''_w) \tag{6}$$

2.4. Objective

We use the commonly used cross-entropy loss for emotion classification and our objective is

$$L = L_{ce}(y - \hat{y}) \tag{7}$$

3. EXPERIMENT

Our proposed method is validated on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [19] dataset. In this section, we firstly introduce the dataset processing and audio sources used. Then we describe our experimental setup and the used validation strategy.

3.1. Datasets

IEMOCAP is a widely used emotion recognition dataset, recorded from ten different actors with audio, video, transcriptions and motion-capture information. Following others' work [12, 7, 5], we merge "happy" and "excited" into the category of "happy" and we consider the 5531 acoustic utterances from 4 emotions, angry, sad, happy and neutral. In order to more accurately evaluate the performance of the model, we test our model with the 5-fold leave-one-session-out and the 10-fold leave-one-speaker-out cross-validation strategy to generate the speaker-independent results. Also, we use the commonly used weighted accuracy (WA) and the unweighted accuracy (UA) as the evaluation metrics.

3.2. Experimental Setup

The used raw audio signals are sampled at 16 kHz. We spilt each audio utterance into several segments with a length of 3 seconds. When a segment is less than 3 seconds, a padding operation with 0 will be applied to this segment to keep the same length. The final prediction result of an audio utterance will be decided by all split segments from this utterance.

To make full use of different levels of speech information, we use three kinds of acoustic information in this SER task, MFCC, spectrogram and W2E. MFCC is a 40-dimension HTK-style Mel frequencies feature that taking into account the human auditory characteristics. It is extracted from the raw audio segments with librosa library [20]. Spectrogram and the W2E are the deep features of audio signals. For spectrogram, a series of 40-ms Hamming windows with a hop length of 10 ms is applied and here we treat each windowed block as a frame. Each frame is transformed into a frequency domain with the Discrete Fourier Transform (DFT) of length 800. The first 200 DFT points are used as input spectrogram features. We finally get a spectrogram image with a size of 300*200 for each audio segment. Like the multimodal emotion recognition method [21], W2E are obtained from the pre-trained transformer-based wav2vec2 network. It is the reflection of the deep feature of speech in the time domain.

This SER system is implemented in PyTorch. The optimizer for the model is AdamW with a learning rate of 1e-5. The training batch size is 64 and we set the early stopping setting as 8 epochs. Our code will be available on Github[1].

4. RESULTS AND ANALYSIS

In this section, we present the model performance and design an ablation study to evaluate the influence of different inputs and used modules. We also visualize the extracted features of our model with t-distributed stochastic neighbour embedding (t-SNE) and the final normalized confusion matrix.

[1] https://github.com/Vincent-ZHQ/CA-MSER

Table 2. Ablation study on the proposed model.

Model	WA	UA
MFCC	57.60	58.09
Spectrogram	62.13	62.25
W2E	64.03	65.67
MFCC+W2E (w/o co-att)	64.62	65.93
Spectrogram+W2E (w/o co-att)	66.20	67.22
MFCC+Spectrogram+W2E (w/o co-att)	67.22	67.81
W2E (w/ co-att)	67.55	68.65
MFCC+W2E (w/ co-att)	69.11	70.30
Spectrogram+W2E (w/ co-att)	70.05	71.30
MFCC+Spectrogram+W2E (w/ co-att)	**71.64**	**72.70**

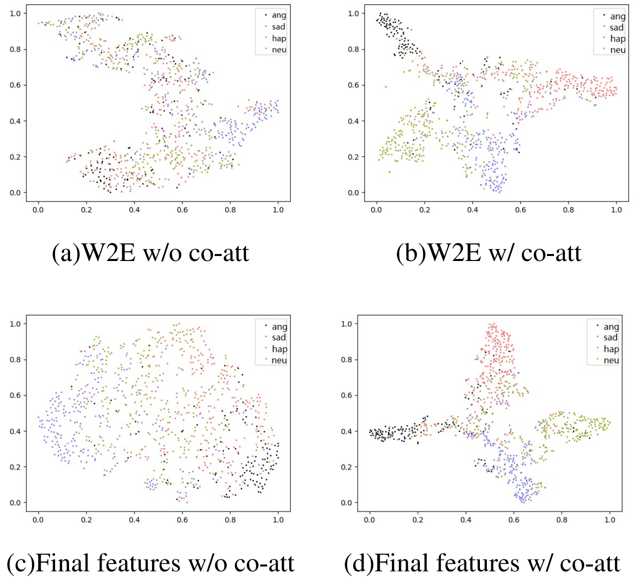
4.1. Results and Comparison

As shown in Table 1, our proposed method could achieve the best performance of 69.80% and 71.05% in terms of UA and WA for the leave-one-session-out validation strategy. And for the leave-one-speaker-out validation strategy, this method could also achieve the highest UA with a value of 72.70%. At the same time, its performance in WA is also competitive with a very similar result of 71.64% compared with UA on this unbalanced IEMOCAP dataset.

4.2. Ablation Study

Our proposed method utilizes the multiple levels of acoustic information, which contains the time domain and frequency domain. Table 2 shows the ablation study of model performance with different combinations of acoustic information. The first three rows are the emotion recognition results with only one level of acoustic information: MFCC, spectrogram and W2E. W2E provides better performance than the others for the final emotion recognition. The next three rows summarize the results from the combination of different features with W2E. The last four rows present the results of different combination features with the weighted W2E information after co-attention. The combination of multiple acoustic information and the proposed co-attention module are observed to contribute a lot to improve the whole model's performance.

The ablation study also shows the effectiveness of the proposed co-attention mechanism. From the last two rows of Table 2, the co-attention mechanism further optimizes the fused data and performs better than the direct concatenating operation with 4.42% and 4.89% improvement on WA and UA, respectively. As shown in Fig.2, the t-SNE visualization of the weighted W2E and final combined features after co-attention present a much more clear classification boundary when compared with the results of the unweighted W2E and final combined features without co-attention. From Fig. 3, we also observe that the final classification results of the model with co-attention are much better than the model without co-attention from the final normalized confusion matrix.



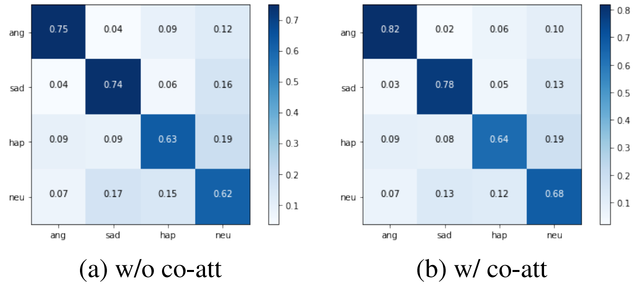
(a)W2E w/o co-att

(b)W2E w/ co-att

(c)Final features w/o co-att

(d)Final features w/ co-att

Fig. 2. The t-SNE visualization of feature distribution. (a) and (b) are the final extracted W2Es in the model trained with multi-level acoustic information without and with the proposed co-attention. (c) and (d) are the final combined features without and with the proposed co-attention



(a) w/o co-att

(b) w/ co-att

Fig. 3. The normalized confusion matrix for the final speech emotion recognition without and with the proposed co-attention module.

5. CONCLUSION

This paper proposes a co-attention-based SER system utilizing multi-level acoustic information. By designing different encoders, this model could get feature-specific information from the raw audio signals and enables complementary acoustic information for the SER problem. Also, this method introduces a co-attention based fusion method for getting weighted wav2vec2 embeddings and combining the final features. The experiments on the IEMOCAP dataset show that our proposed method achieves competitive performance with different speaker-independent cross-validation methods. In the future, we would like to combine the knowledge from different languages or datasets to improve the final performance.

6. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[3] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3020–3024.

[4] E. Guizzo, T. Weyde, and J. B. Leveson, "Multi-time-scale convolution for emotion recognition from speech audio signals," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6489–6493.

[5] Q. Cao, M. Hou, B. Chen, Z. Zhang, and G. Lu, "Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6334–6338.

[6] Y. Gao, J. Liu, L. Wang, and J. Dang, "Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6314–6318.

[7] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6269–6273.

[8] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4275–4279.

[9] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3362–3366.

[10] C. Li, B. Chen, Z. Zhao, N. Cummins, and B. W. Schuller, "Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1240–1244.

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[12] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6304–6308.

[13] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition." in *Interspeech*, 2019, pp. 2578–2582.

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

[15] 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.

[16] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," 2019.

[17] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu *et al.*, "Speech emotion recognition using capsule networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6695–6699.

[18] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L.-R. Dai, "An attention pooling based representation learning method for speech emotion recognition," 2018.

[19] 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," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[21] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech*, 2021.