ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096709

EXPLORING COMPLEMENTARY FEATURES IN MULTI-MODAL SPEECH EMOTION RECOGNITION

Suzhen Wang[1], *Yifeng Ma*[2], *Yu Ding*[1]*

[1]Virtual Human Group, Netease Fuxi AI Lab, China; [2]Tsinghua University, China
{wangsuzhen, dingyu01}@corp.netease.com, mayf18@mails.tsinghua.edu.cn

ABSTRACT

Speech emotion recognition (SER) is of great importance in human-computer interaction. Recent research has demonstrated that self-supervised learned acoustic and linguistic features are helpful in this task. However, few works have fully exploit the advantages of the pre-trained features in SER. The primary challenge is how to effectively extract the complementary emotional information implied in the pre-trained features of the respective modality. To tackle this challenge, we propose a novel modality-sensitive multimodal speech emotion recognition framework. In a nutshell, we aim to exploit the typical emotion features in each modality and then fuse the complementary emotional information for classification. Specifically, we first utilize the parallel uni-modal encoders to refine the emotion-related information from the pre-trained features of each modality. For better fusion of the multimodal features, we develop a group of learnable emotion query tokens to gather the emotional information from the refined acoustic and linguistic features with the cross-attention mechanism in the transformer decoder. Observing the modality bias problem in multimodal methods, we introduce the random modality masking training strategy to maximize the utilization of the emotional information in each modality and mitigate this problem. We evaluate our method on the widely used IEMOCAP dataset and achieve 1.1% and 0.9% improvements on the unweighted accuracy and weighted accuracy, respectively. Extensive experiments demonstrate the effectiveness of the proposed method.

Index Terms— Speech Emotion Recognition, Multimodal Fusion, Transformer

1. INTRODUCTION

With the rise of the metaverse, *Speech Emotion Recognition* (SER) plays a vital role in human-computer interaction (HCI) [1] and affective computing [2]. With the help of SER, the machine can perceive emotions in the context of conversations and then respond to humans appropriately.

When we speak, the content of our message and the tone of our voice complement each other to convey our feelings. Due to this nature, multimodal SER has attracted increased attention in recent years. *Yoon et al.* [3] propose a novel multimodal dual recurrent encoder model to combine the audio signal and text feature for SER. *Gu et al.* [4] design a deep multimodal architecture for SER by aligning text and audio at the word level and applying hierarchical attention to textual and acoustic features. *Peri et al.* [5] introduce video information and set up a multitask learning for emotion recognition.

In this paper, we aim to recognize the emotion from acoustic and linguistic information. With the great success of large-scale self-supervised pre-trained models for natural language processing [6, 7] and speech representation extraction [8, 9], some works [10, 11] opt to utilize the self-supervised learned features and achieve better performance in SER. Recently, MPNet [7] and HuBERT [9] are the widely used pre-trained models and have been shown to be robust in many tasks. Thus, we employ them to extract text and audio features in this paper.

Inspired by the powerful pre-trained features, a series of methods [12, 13, 14, 11, 15] with various multimodal fusion strategies have been proposed for SER. However, few works focused on exploring the complementary emotional information implied in the pre-trained features of the respective modality. In practice, we find that multimodal methods are apt to classify emotions mainly relying on linguistic information while making light of audio signals. One assumption is that the emotional information lies in the plain sight of the pre-trained linguistic features, while it lies in the deep space of the audio features; therefore, the network may rapidly converge toward the text direction. However, audio information should be at least as significant as text, if not more so. For example, different tones may imply different emotions when speaking the same utterance.

To address the above issues, we propose a novel **M**odality-**S**ensitive **M**ultimodal **S**peech **E**motion **R**ecognition framework (**MSMSER**). To be specific, we develop a modality interaction transformer [16] (MIT), which utilizes a group of trainable emotion query tokens to retrieve emotional information from the concatenated text and audio features. Benefiting from the design of MIT, we introduce a Random Modality Masking (RMM) strategy by randomly masking the audio features or text features during training to compel

* Corresponding author.

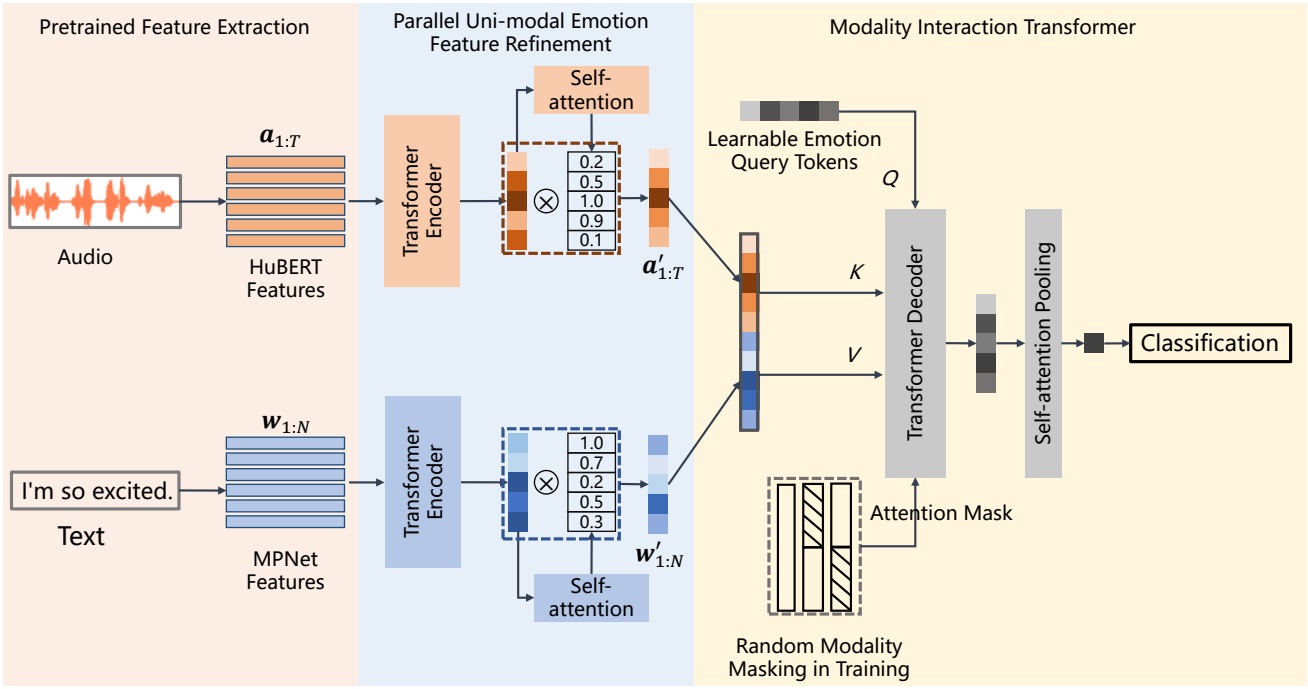

Fig. 1. Overview of the proposed modality-sensitive multimodal speech emotion recognition framework. We utilize the pre-trained HuBERT [9] and MPNet [7] to extract the acoustic and linguistic features, respectively. In order to refine the emotional information in the pre-trained embeddings, we develop two parallel uni-modal emotion feature refinement encoders followed by self-attention modules. Then we propose a modality interaction transformer to integrate the emotion features from two modalities through the learnable emotion query tokens. Moreover, we design a random modality masking training strategy by randomly masking the audio or text embeddings in training to compel the models to fully perceive the emotional information of each modality.

the model to fully perceive the powerful emotion features in each modality. Thanks to this mechanism, our method greatly mitigates the modality bias problem. Extensive experiments on the widely used IEMOCAP dataset demonstrate that the proposed method exceeds the state-of-the-art approaches.

2. PROPOSED METHOD

Given the text and audio signal of one speech, our method is able to infer the emotion label. The overview of the proposed framework is shown in Figure 1. In this work, we employ the well-known self-supervised learning models HuBERT [9] and MPNet [7] to extract the pre-trained acoustic and linguistic features, respectively. We introduce the parallel uni-modal emotion feature refinement encoders in Section 2.1 to fully encode the emotion-related information in the self-supervised features. Then in Section 2.2, we develop a novel multi-modal fusion approach to capture the complementary emotion features from the acoustic and linguistic embeddings. To overcome the problem of modality bias, we propose a random modality masking strategy to make our model sensitive to emotional information in each modality. This strategy is presented in Section 2.3.

2.1. Uni-modal Emotion Feature Refinement

Since neither HuBERT nor MPNet is pre-trained for emotion recognition, we introduce two parallel uni-modal transformer encoders to exploit the emotional information in the self-supervised features. Given the pre-trained frame-level audio features $\boldsymbol{a}_{1:T} \in \mathbb{R}^{T\times d_a}$ and word-level text features $\boldsymbol{w}_{1:N} \in \mathbb{R}^{N\times d_w}$, we obtain the emotion-related features after encoding. Afterwards, the encoded features are further refined with a self-attention module to highlight the typical emotional frames in the respective modalities. The self-attention module consists of a two-layer Multilayer Perceptron (MLP) and a Sigmoid to obtain the scales for each frame. The encoded features are then scaled to obtain the final text and audio features, $\boldsymbol{a}'_{1:T} \in \mathbb{R}^{T\times d_e}$ and $\boldsymbol{w}'_{1:N} \in \mathbb{R}^{N\times d_e}$. Note that $\boldsymbol{a}_{1:T}$ and $\boldsymbol{w}_{1:N}$ are transformed to the same dimension d_e before encoding.

2.2. Modality Interaction Transformer

Different from the multimodal fusion strategy in previous works, we develop a modality interaction transformer (MIT) to better integrate the emotional features from different

sources. To be specific, we employ a group of trainable emotion query tokens to aggregate emotional information of the concatenated $\boldsymbol{a}'_{1:T}$ and $\boldsymbol{w}'_{1:N}$ by the attention mechanism. The query tokens $q \in \mathbb{R}^{L \times d_e}$ are used as the query ($\boldsymbol{Q}$), and the assembled multimodal features are used as the key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) in the transformer decoder. L is the number of query tokens. Before concatenation, the text features and audio features are appended with position encodings and corresponding trainable type codes. The position encodings help the model perceive contextual information, and the trainable type codes make the emotion query tokens sensitive to the features of different modalities. Since we calculate the cross-attentions between each query token and all the acoustic and linguistic features, the query tokens can be viewed as the intermediary which helps the emotional information in two modalities interact with each other. After cross-attention, the output L emotion query tokens contain the complementary emotional features. Then we employ a self-attention pooling layer [14] to merge L tokens into one and send it to a 2-layer MLP for classification.

2.3. Random Modality Masking

In practice, we observe that the model tends to recognize emotions primarily using the emotion features in the text. This problem is demonstrated in Section 4.3. To mitigate this problem, we introduce the random modality masking strategy (RMM) when training the models. Taking advantage of the structure of MIT (Section 2.2), we can easily mask any modality feature by setting the corresponding attention mask to zero. In the early stages of training, we make our model pay more attention to the uni-modal features by randomly masking the audio or text embeddings. Then in the later stage, the model focuses on obtaining the complementary multimodal emotional features. Specifically, we introduce a hyperparameter p, which denotes the probability that the model applies the RMM in each training sample. RMM masks the text features with a probability 0.6 and the audio features with a probability 0.4, respectively. In our experiments, p starts at 0.8 at and decays to 0 as the training continues.

3. EXPERIMENTS

3.1. Dataset

Following the most previous work on SER, we use the IEMOCAP [17] dataset. This dataset contains five sessions, and each session is recorded during a conversation between one male and one female speaker. To stay consistent with previous work [18, 15, 11], we consider the 5531 acoustic utterances of 4 emotions, neutral, happy (happy & excited), sad and angry. For more accurate evaluations, we conduct our experiments with the 5-fold leave-one-session-out cross-validation. Also, we use the commonly used weighted accuracy (WA) and the unweighted (UA) as the evaluation metrics.

Table 1. Comparison with other methods on IEMOCAP (%).

Method	UA (%)	WA (%)
CMA [12]	72.8	-
GBAN [14]	70.1	72.4
STSER [19]	72.1	71.1
CME [20]	73.5	72.7
KS-Transformer [11]	75.3	74.3
MSMSER (Ours)	**76.4**	**75.2**

3.2. Implementation Details

Our framework is implemented by PyTorch. The pre-trained models of HuBERT and MPNet are available online[1]. We employ Adam optimizer for training with a learning rate of 2×10^{-5}. The hidden size of the 8-head transformer encoder and decoder is set to 768. Our models are trained for 100 epochs, and the random modality masking probability p is 0.8 and linearly decays to 0 at the 70th epoch. The batch size is 8. Our code will be available soon.

4. RESULTS AND ANALYSIS

To verify the effectiveness of the proposed framework, we carry out 4 groups of experiments on IEMOCAP. In Section 4.1, we compare our method with the recent multimodal SER approaches. In Section 4.2, 4.3 and 4.4, we perform a series of ablation studies to prove the effectiveness of each component in our method.

4.1. Results and Comparison

We first compare our methods with other multimodal (text + audio) SER methods on the IEMOCAP, and the comparison result are listed in Table 1. The baseline methods include CMA [12], GBAN [14], STSER [19], CME [20], KS-Transformer [11] and all the quantitative results are collected from their papers. It can be observed that our method outperforms the state-of-the-art by 1.1% and 0.9% in terms of unweighted accuracy and weighted accuracy, respectively.

4.2. Ablation Study on the Subcomponents

To evaluate the effectiveness of each component in our framework, we first conduct an ablation study with 7 variants: (1) Only use the pre-trained text feature (**Text**), (2) Only use the pre-trained audio feature (**Audio**), (3) remove the uni-modal encoder in Section 2.1 (**w/o Encoder**), (4) remove the self-attention module in Section 2.1 (**w/o SelfAtt**), (5) remove the modality interaction transformer with concatenation fusion model (**w/o MIT**), (6) and our full model (**Full**). The results are reported in Table 2. As shown, all multimodal meth-

[1] https://github.com/facebookresearch/fairseq

Table 2. Results of ablation study on each component of the proposed framework.

Method	UA (%)	WA (%)
Text	66.5	64.7
Audio	64.9	63.2
w/o Encoder	71.9	70.9
w/o SelfAtt	72.5	71.4
w/o MIT	67.2	65.1
Full	**76.4**	**75.2**

Table 3. Results of ablation study on the random modality masking strategy.

Method	UA (%)	WA (%)
w/o RMM	73.3	71.7
w/o RMM (only text)	63.2	61.2
w/o RMM (only audio)	55.5	55.2
Full	**76.4**	**75.2**
Full (only text)	65.8	64.4
Full (only audio)	63.4	62.9

ods surpass the uni-modal methods (**Text** or **Audio**). Compared with the **Full** model, the performance of **w/o Encoder** drops dramatically, demonstrating that the two parallel uni-modal encoders help exploit the emotional information. Furthermore, the comparisons between **w/o Encoder** and **w/o SelfAtt** imply the effectiveness of the self-attention module in each uni-modal encoder. The performance drops dramatically by replacing the modality interaction transformer with the simple concatenation fusion method, which shows that it is difficult for the commonly used fusion strategy to extract the complementary information. Once the modality interaction transformer is ablated, the performance drops dramatically. Therefore, the design of MIT significantly improves aggregating the emotional information in each modality.

4.3. Evaluation of RMM

We then perform another group of ablation studies to evaluate the effectiveness of the random modality masking strategy. We train the models with and without RMM, denoted as **Full** and **w/o RMM**, respectively. Then we evaluate both models by masking the audio or text in inference stage. The results are shown in Table 3. We can observe that the performance of masking the text features (**w/o RMM (only audio)**) is much lower than that of masking the audio features (**w/o RMM (only audio)**) when testing the trained models without RMM. It illustrates that the models tend to use the text features for classification. While it shows the approximate results in the same test settings when using the models trained by RMM. Furthermore, the results of **Full (only text)** and **Full (only audio)** are close to the results of the uni-modal (**Text** and **Audio**) in Table 2. It proves that RMM helps our model take full advantage of the emotional information in each modality and mitigates the modality bias problem.

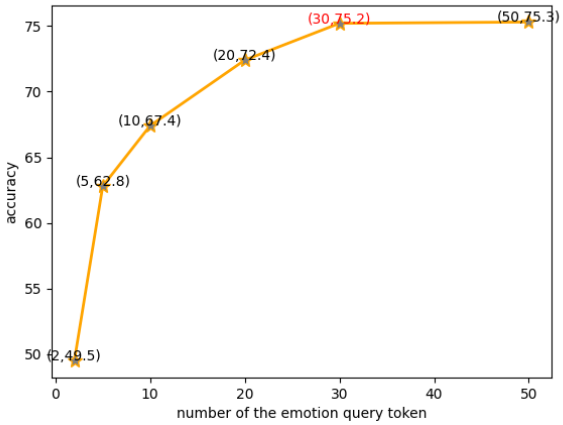

Fig. 2. Visualization of the correlation between the performance and the number of the emotion query token.

4.4. Analysis of the Emotion Query Token

We conduct another group of experiments to explore the appropriate number of the emotion query token introduced in Section 2.2. The query tokens are trained to perceive the emotional information in each modality. In our experiments, we set the number L of the query token at 2, 5, 10, 20, 30, and 50. The effect of the number of the emotion query token is shown in Figure 2. As can be seen, the performance increases as the L grows. We choose $L = 30$ in our final model for the tradeoff.

5. CONCLUSION

In this paper, we propose a novel modality-sensitive multimodal speech emotion recognition framework (MSMSER). We first use two parallel uni-modal encoders with a self-attention module to refine the pre-trained audio and text features. Then we utilize a group of learnable emotion query tokens to incorporate the emotional information from multimodal features with a transformer decoder. In addition, we propose to use the random modality masking strategy to overcome the modality bias problem in the multimodal task. Our method exhibits a new perspective for effectively fusing the multimodal features. Extensive experiments on the widely used IEMOCAP dataset prove the superiority of the proposed framework. We plan to combine more modalities to improve SER performance in our future work.

6. REFERENCES

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] Carlos Busso, Murtaza Bulut, Shrikanth Narayanan, J Gratch, and S Marsella, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2013.

[3] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[4] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2018, vol. 2018, p. 2225.

[5] Raghuveer Peri, Srinivas Parthasarathy, Charles Bradshaw, and Shiva Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6344–6348.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NIPS*, vol. 33, pp. 12449–12460, 2020.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, "Speech emotion recognition with multi-task learning," in *Interspeech*, 2021.

[11] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP*. IEEE, 2022, pp. 6897–6901.

[12] DN Krishna and Ankita Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks.," in *Interspeech*, 2020, pp. 4243–4247.

[13] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[14] Ming Chen and Xudong Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition.," in *INTERSPEECH*, 2020, pp. 374–378.

[15] Soumya Dutta and Sriram Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6917–6921.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[18] Qi Cao, Mixiao Hou, Bingzhi Chen, Zheng Zhang, and Guangming Lu, "Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition," in *ICASSP 2021*. IEEE, 2021, pp. 6334–6338.

[19] Pengfei Liu, Kun Li, and Helen Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," *arXiv preprint arXiv:2201.06309*, 2022.

[20] Hang Li, Wenbiao Ding, Zhongqin Wu, and Zitao Liu, "Learning fine-grained cross modality excitement for speech emotion recognition," *arXiv preprint arXiv:2010.12733*, 2020.