LETTER

A Two-Stage Attention Based Modality Fusion Framework for Multi-Modal Speech Emotion Recognition*

Dongni HU[†,††], **Chengxin CHEN**[†,††], **Pengyuan ZHANG**[†,††a)], **Junfeng LI**[†,††], **Yonghong YAN**[†,††], *Nonmembers*, *and* **Qingwei ZHAO**[†], *Member*

SUMMARY Recently, automated recognition and analysis of human emotion has attracted increasing attention from multidisciplinary communities. However, it is challenging to utilize the emotional information simultaneously from multiple modalities. Previous studies have explored different fusion methods, but they mainly focused on either inter-modality interaction or intra-modality interaction. In this letter, we propose a novel two-stage fusion strategy named modality attention flow (MAF) to model the intra- and inter-modality interactions simultaneously in a unified end-to-end framework. Experimental results show that the proposed approach outperforms the widely used late fusion methods, and achieves even better performance when the number of stacked MAF blocks increases.

key words: *speech emotion recognition, multi-modal fusion, attention mechanism, end-to-end*

1. Introduction

Speech emotion recognition (SER) has received increasing attention because of its application in human-computer interaction (HCI), mental health analysis and improvement of customer service. However, it remains challenging to accurately recognize emotion from speech due to its subjectivity and ambiguity [1]. Since humans express emotion via multiple modes, such as acoustic,textual and visual, it is expected that fusing the information from different modalities could outperform uni-model approaches [2]. In this paper, we focus on how to utilize both acoustic and textual information from speech to further improve the accuracy of SER.

There are two common ways for modality fusion of SER. In the late fusion, the predictions of uni-modal branches are fused (concatenated or multiplied) to make a final prediction. Tripathi et al. [3] built a late fusion network based on the cLSTM block and achieved state-of-the-art performance. The late fusion is effective at modelling intra-modality interactions, *i.e.*, frame-to-frame relations and word-to-word relations, but poor at inter-modality interactions, *i.e.*, frame-to-word relations and word-to-frame relations. In contrast, the early fusion could model interactions across modalities at raw features stage. Georgiou et al. [4] concatenated features from different modalitiy at various levels and used multi-layer perceptron for emotion prediction. Generally speaking, concatenation based early fusion methods do not outperform the late fusion methods in SER [5]. Many other works [6]–[8] considering multimodal SER aimed at improving recognition accuracy by adopting tons of hand-crafted features or pre-trained unimodal branches before late fusion. Only a few works proposed different fusion mechanism considering both intra- and inter-modality fusion. Yoon et al. [9] fused encoded features with multi-hop attention, where one modality is used to direct attention for the other mode. Pan et al. [10] utilized a multi-modal attention sub-network to model inter-modality interactions before late fusion. Since intra-modality interactions focus on emotionally salient parts within each modality and inter-modality interactions imply the latent alignments, we argue that both intra- and inter-modality interactions are important.

Inspired by the concept of attention flow proposed by Peng et al. [11], we proposed a novel two-stage attention-based modality fusion framework using acoustic and textual cues for SER. Different from the interweaved intra- and inter-modality attention blocks in [11], we argue that intra- and inter-modality interactions share equal priority. Consequently, we model the elementary intra- and inter-modality interactions simultaneously in the first stage, and explore the advanced interactions based on the elementary ones in the second stage. To our best knowledge, it is the first time the intra- and inter-modality interactions are modelled in a unified end-to-end framework for SER. We evaluate the approach on the well benchmarked IEMOCAP dataset [12] and the experimental results show that the proposed approach achieves state-of-the-art performance. We further evaluate the approach on a much larger multi-modal emotional dataset used in Multimodal Emotion Recognition Competition 2020 (MERC2020) and the experimental results show considerable improvement of performance over the official baseline.

2. Framework

2.1 Overview

The whole pipeline of the proposed approach is illustrated in Fig. 1. The extracted acoustic and textual features are

Manuscript received January 6, 2021.
Manuscript revised April 15, 2021.
Manuscript publicized April 30, 2021.

[†] The authors are with Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100864, China.

[††] The authors are with University of Chinese Academy of Sciences, Beijing, 100049, China.

[*] This work is partially supported by the National Key Research and Development Program (No. 2020AAA0108002).

a) E-mail: zhangpengyuan@hccl.ioa.ac.cn

DOI: 10.1587/transinf.2021EDL8002

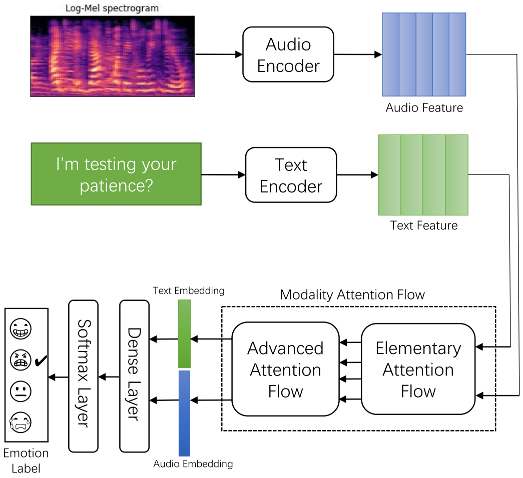

Fig. 1 Framework of the proposed approach.

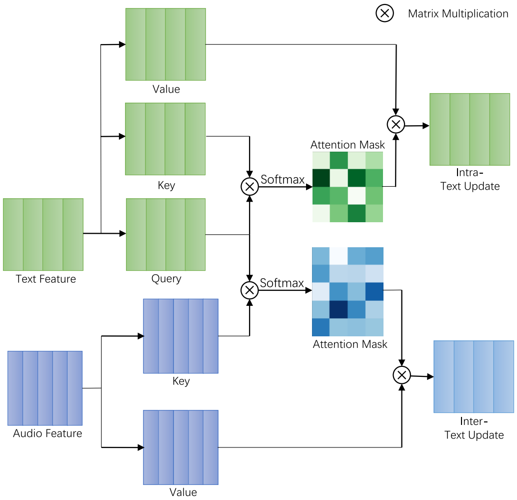

Fig. 2 The structure of EAF on textual modality.

first sent into the Elementary Attention Flow (EAF) module to learn the cross-modal interactions between the speech frames and the words, as well as the relationships within each modality. In order to further model emotional salient parts of each modality and the latent alignments between modalities, the Advanced Attention Flow (AAF) module is proposed. The detailed implementation of the two modules, which build up the Modality Attention Flow (MAF), will be introduced in the following sections. Subsequently, the outputs are transformed into 1-dimensional vectors respectively with mean pooling over time. The obtained embeddings are finally concatenated before sent into the emotion classifier, which consists of multiple dense layers and a softmax layer.

2.2 Bidirectional Recurrent Encoder

In the data preprocessing stage, 128-dimensional log-mel spectrograms of speech (denoted as S) with 40 ms frame length and 10 ms frame shift are extracted, and the corresponding transcripts are converted to a sequence of 300-dimensional vectors using GLoVe word embeddings (denoted as E) [13]. To model the sequential property of the acoustic and textual signals, we adopt Bidirectional Long Short-Term Memory (BLSTM) network as encoders, and the forward/backward hidden states are concatenated as outputs. The obtained acoustic feature A and textual feature T could be denoted as

$$A = \mathrm{BLSTM}(S; \theta_{audio}) \tag{1}$$

$$T = \mathrm{BLSTM}(E; \theta_{text}) \tag{2}$$

where the parameters θ_{audio} and θ_{text} are both learned from scratch and updated together during the training. In order to compute in parallel, A and T are padded and truncated to the same length respectively (denoted as u and v).

2.3 Elementary Attention Flow

In the Elementary Attention Flow module, each acoustic and textual features are first transformed into query, key and value features with fully-connected layer. The transformed acoustic features are denoted as $A_Q, A_K, A_V \in R^{u \times dim}$, while the transformed textual features are denoted as $T_Q, T_K, T_V \in R^{v \times dim}$, where dim represents the common dimension of transformed features from both modalities.

Figure 2 demonstrates the calculating process on the textual modality. The intra-modality matrices and inter-modality matrices, which capture the importance between words and information from frames to words respectively, could be defined as,

$$\mathrm{IntraEAF}_{T \to T} = \mathrm{SoftMax}(\frac{T_Q T_K^T}{\sqrt{dim}}) \tag{3}$$

$$\mathrm{InterEAF}_{A \to T} = \mathrm{SoftMax}(\frac{T_Q A_K^T}{\sqrt{dim}}) \tag{4}$$

Subsequently, the value features of the two modalities are multiplied by the derived attention masks to update textual-relevant features,

$$IntraT_{update} = \mathrm{IntraEAF}_{T \to T} \times T_V \tag{5}$$

$$InterT_{update} = \mathrm{InterEAF}_{A \to T} \times A_V \tag{6}$$

The updated textual features are denoted as $IntraT_{update} \in R^{v \times dim}$ and $InterT_{update} \in R^{v \times dim}$ respectively. After concatenating the updated features with original textual features, a fully connected layer is utilized to transform the concatenated features into output features,

$$IntraT = \mathrm{Linear}([T, IntraT_{update}]^T; \theta_{IntraT}) \tag{7}$$

$$InterT = \mathrm{Linear}([T, InterT_{update}]^T; \theta_{InterT}) \tag{8}$$

Simultaneously, $IntraA$ and $InterA$ are calculated in a symmetrical way, which encode intra- and inter-modal relations with query from acoustic modality.

2.4 Advanced Attention Flow

With the outputs of EAF ($IntraA, InterA, IntraT, InterT$),

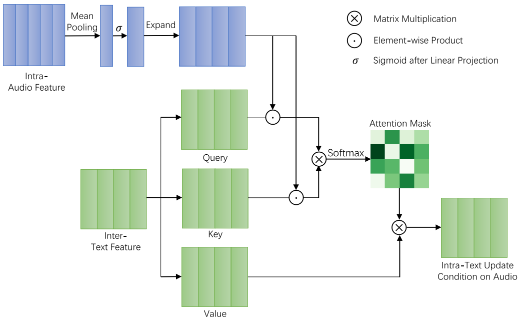

Fig. 3 The structure of AAF on textual modality.

we can go a step further to model higher-order attention flow in the Advanced Attention Flow module, which means we can calculate inter-modal relations based on intra-modal relations, and vice versa. Figure 2 demonstrates the calculating process on the textual modality. Since *InterT* has encoded the inter-modal relation from frames to words, we propose to calculate the intra-modal relation within words based on *InterT*. Different from the typical self-attention paradigm, we use the information from *IntraA*, which has encoded the intra-modal relation within speech frames. The motivation is that the relations between different word pairs should be weighted differently according to speech. In this way, the second-order attention flow is accomplished.

Firstly, *IntraA* is transformed into a channel-wise conditioning vector to distill the information from acoustic modality,

$$G_{A\rightarrow T} = \sigma(\text{Linear}(\text{Mean\_Pool}(IntraA)); \theta_{TP}) \tag{9}$$

where σ denotes sigmoid non-linearity function. The query and key features of *InterA* are then modulated by G_{T2A},

$$IntraT_Q^G = (1 + G_{A\rightarrow T}) \circ IntraT_Q \tag{10}$$

$$InterT_K^G = (1 + G_{A\rightarrow T}) \circ IntraT_K \tag{11}$$

where $\circ$ denotes element-wise multiplication. Similar to Eq. (3), (5), (7), we can obtain $IntraT^G$. Simultaneously, $IntraA^G$ is calculated in a symmetrical way. They are the final outputs of the MAF block.

3. Experiments

3.1 Dataset

We use two datasets to evaluate the proposed approach. The IEMOCAP dataset is a well benchmarked dataset in English, which contains approximately 12 hours of audio-visual data from 10 actors. The dataset contains 5 sessions and each session is performed by one female and male actor in scripted and improvised scenarios, and the ground truth transcripts of speech are provided. To be consistent with the previous works, we merge the utterances labeled 'excited' into the 'happy' class, and the distribution of utterances used in the experiments is {happy: 1636, sad: 1084, angry: 1103, neutral: 1708}. Another dataset named MERC2020 is provided by KAIST in Multimodal Emotion Recognition Competition 2020. It's a much larger multi-modal emotional dataset in Korean, with approximately 78 hours of audio-visual data from 95 people, including 7 balanced emotional labels named {neutral, happy, angry, fear, disgust, surprise, sad}.

3.2 Reference Baselines

Three baselines are constructed to make a comparison with the performance of the proposed BLSTM-MAF model.
Speech-only BLSTM (BLSTM-A): The BLSTM-A baseline receives acoustic features only, which are passed through the audio encoder (two BLSTM layers with a dropout rate of 0.5) and a self-attention module for prediction.
Transcript-only BLSTM (BLSTM-T): The BLSTM-T baseline is similar to BLSTM-A except that it receives textual features only.
Bimodal BLSTM with late fusion (BLSTM-LF): The BLSTM-LF baseline has a hierarchical structure. The lower level consists of two uni-modal networks BLSTM-A and BLSTM-T. At the higher level, the embeddings of the two uni-modal networks are concatenated for the final prediction.

3.3 Experimental Setup

The hyper-parameters for the two modalities are the same in all experiments. The max length of audio is set to 6s and the max length of transcript is set to 40 words. The batch size is set to 64, and the max training epochs are set to 50 with early stopping mechanism. Adam optimizer with a learning rate of 1e-3 is applied to optimize the model parameters. For IEMOCAP dataset, we use Leave One Speaker Out (LOSO) 10-fold cross validation. For MERC2020 dataset, the train and val split set are provided officially, and the distribution of it is {train: 44370, val: 5386}.

3.4 Performance Evaluation

To be consistent with the previous works, we evaluate the performance on IEMOCAP dataset using unweighted average recall (UAR) and weighted average recall (WAR). While Accuracy (Acc) is adopted for evaluation on MERC2020 dataset to be compared with official baseline. All the experiments are repeated with random initialization for five times to reduce deviation, and the UAR, WAR, Acc reported in this letter are the average results.

The experimental results are shown in Table 1. We can discover that on both datasets, BLSTM-LF achieves a significant improvement on all evaluation metrics compared with uni-modal networks, suggesting that the textual cues do complement the information of speech in emotion recognition. The proposed BLSTM-MAF framework outperforms BLSTM-LF, suggesting that both modality-specific and cross-modal interactions are important in emotion recognition.

Since the inputs(A, T) and oupts$(IntraA^G, IntraT^G)$ of

Table 1 Performance comparison of different models.

Model	IEMOCAP		MERC2020
	UAR(%)	WAR(%)	Acc(%)
BLSTM-A	61.44	58.41	39.65
BLSTM-T	61.55	61.66	44.39
BLSTM-LF	70.54	69.26	50.29
BLSTM-MAF	71.79	69.98	50.89
BLSTM-MAF2	**72.86**	**71.61**	51.92
BLSTM-MAF4	70.94	70.16	**53.26**

Table 2 Comparison with previous state-of-the-art methods on IEMOCAP dataset.

Method	Acoustic feature	UAR(%)	WAR(%)
MDRE [6]	FBank+MFCC	-	71.8
Att-align [8]	FBank+MFCC	70.9	**72.5**
STSER [7]	FBank	72.05	71.06
cLSTM-MMA [10]	IS13-ComParE [14]	-	71.66
Proposed(MAF2)	FBank	**72.86**	71.61

Table 3 Comparison with official baseline on MERC2020 dataset.

Method	Modality	Acc(%)
official	Acoustic	39.3
official	Textual	45.2
official	Visual	29.7
official(Late fusion)	Acoustic+Textual+Visual	52.6
Proposed(MAF4)	Acoustic+Textual	**53.26**

MAF block are of the same shape, we can stack the MAF block multiple times to achieve even deeper modality fusion. Ablation study on the number of stacked MAF blocks is conducted. On IEMOCAP dataset, BLSTM-MAF2 outperforms BLSTM-MAF, suggesting that higher order interactions within and cross modalities are learned when the number of stacked MAF blocks increases. However, the performance is even worse than BLSTM-LF when the number increases to four. It's possible that the utilized utterances are not enough to train such a complicated network, which leads to overfitting. The experimental results on MERC2020 prove this assumption. On MERC2020, which is seven times as large as IEMOCAP, BLSTM-MAF4 achieves the highest Acc in all the models.

In Table 2, We compare the performance of BLSTM-MAF2 with other multi-modal approaches using acoustic and textual cues on IEMOCAP. Our approach achieves the highest 72.86% UAR and 71.61% WAR. Although the WAR of [8], [10] is higher, they use more complex manual designed acoustic feature, which needs more extra professional knowledge. In Table 3, we compare the performance of BLSTM-MAF4 with official baseline on MERC2020. We can discover that our proposed approach can even outperform the official model that use the information from three modalities.

4. Conclusions

In this letter, we proposed a novel two-stage attention-based modality fusion framework using acoustic and textual information for emotion recognition. The elementary intra- and inter-modality relations are first modelled, and the advanced inter-intra and intra-inter attention flows are learned based on the elementary relations. Extensive experiments are conducted on two datasets and the results show the significant improvements over the baselines. The network shows best performance on IEMOCAP with two stacked MAF blocks, achieving the state-of-the-art performance of 72.86% UAR and 71.61% WAR. While on MERC2020, the network with four stacked MAF blocks outperform that with two stacked MAF blocks, achieving 53.26% Acc. We conclude that the number of stacked MAF blocks should be carefully designed according to different size of dataset to achieve best performance. In the future work, we will evaluate the performance using complex encoder structures and incorporate visual modality to develop a more robust emotion recognition framework.

References

[1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell benchmarks and ongoing trends," Commun. ACM, vol.61, no.5, pp.90–99, April 2018.

[2] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion, vol.37, pp.98–125, Sept. 2017.

[3] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," arXiv preprint arXiv:1804.05788, 2019.

[4] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," Proc. Interspeech 2019, pp.1646–1650, 2019.

[5] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, vol.33, no.6, pp.17–25, Nov. -Dec. 2018.

[6] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," 2018 IEEE Spoken Language Technology Workshop (SLT), 2018.

[7] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," Proc. Interspeech, pp.374–378, 2020.

[8] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech[J]," arXiv preprint arXiv:1909.05645, 2019.

[9] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp.2822–2826, 2019.

[10] Z. Pan, Z. Luo., J. Yang, and H. Li, "Multi-modal Attention for Speech Emotion Recognition," arXiv preprint arXiv:2009.04107 (2020).

[11] P. Gao, Z. Jiang, H. You, P. Lu, S.C.H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019.

[12] 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, p.335, 2008.

[13] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," Proc. Conf. Empirical Methods in Natural Language Processing, pp.1532–1543, Oct. 2014.

[14] B.W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," Interspeech 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, Sept. 6-10, 2009, pp.312–315, 2009.