Emotion Recognition With Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information

Lili Guo, *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China*

Longbiao Wang, Jianwu Dang, Yahui Fu, and Jiaxing Liu, *Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China*

Shifei Ding, *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China*

People usually express emotions through paralinguistic and linguistic information in speech. How to effectively integrate linguistic and paralinguistic information for emotion recognition is a challenge. Previous studies have adopted the bidirectional long short-term memory (BLSTM) network to extract acoustic and lexical representations followed by a concatenate layer, and this has become a common method. However, the interaction and influence between different modalities are difficult to promote using simple feature fusion for each sentence. In this article, we propose an implicitly aligned multimodal transformer fusion (IA-MMTF) framework based on acoustic features and text information. This model enables the two modalities to guide and complement each other when learning emotional representations. Thereafter, the weighed fusion is used to control the contributions of different modalities. Thus, we can obtain more complementary emotional representations. Experiments on the interactive emotional dyadic motion capture (IEMOCAP) database and multimodal emotionlines dataset (MELD) show that the proposed method outperforms the baseline BLSTM-based method.

Emotion recognition from speech signals plays an important role in human–machine interactions.[1] Accurately distinguishing emotions can help the machines to understand users' intentions, thus providing better interactivity to enhance user experiences. Therefore, speech emotion recognition has attracted continually increasing attention from researchers due to the rapid growth of human–computer interaction.

When we express emotions through speech, linguistic and paralinguistic information complement each other. Paralinguistic information is embodied by acoustic features, which can be obtained directly from the speech signal, while linguistic information is embodied by phonetic transcribed text. Humans can assign different emotions to neutral texts through paralinguistic information. In this case, the advantage of using acoustic features to recognize emotions is greater than lexical information. On the contrary, lexical information has more advantages. Previous studies have shown promising performance improvements by combining lexical information with acoustic features, demonstrating the potential benefits of acoustic-text structures.[2] Lexical information can be regarded as a semantic supplement for the acoustic feature. Multimodal fusion can utilize the complementarity of emotional information from multiple channels (such as text and facial expression) to improve emotion recognition. However, how to effectively utilize the interaction between acoustic and text modalities to obtain the complementary emotional representation is a challenging problem.

Traditional fusion strategies first extract acoustic and lexical features and subsequently feed the

1070-986X © 2022 IEEE
Digital Object Identifier 10.1109/MMUL.2022.3161411
Date of publication 23 March 2022; date of current version 6 July 2022.

concatenated features into a classifier or shallow-layered fusion models.[3] For extracting acoustic representations, the common features are low-level descriptors (LLD) and their statistical functions.[4,5] For lexical information, bag-of-words models and their refinements were mainly used.[4] Tripathi *et al.*[6] utilized the bidirectional long short-term memory (BLSTM) with attention to realize joint acoustic and lexical information extraction. It used LLD features for acoustic inputs, whereas Glove embeddings are adopted to extract lexical features from words. Li *et al.*[7] also adopted the BLSTM with attention model to automatically learn the best temporal features for emotion recognition. Gu *et al.*[8] proposed a deep multimodal feature fusion framework. In addition, there are some late fusion schemes. For instance, Cho *et al.*[9] used an LSTM network to recognize emotions from acoustic features; meanwhile, a multiresolution convolutional neural network (MCNN) was adopted to detect emotions from word sequences. Thereafter, they fused these two systems to generate the final result. Although these methods have achieved success, they have difficulties in learning mutual relationships among different modalities.

Recently, a more effective no-recurrence transformer model was proposed.[10] It can model information from different time sequences, which is more suitable for modeling acoustic-text emotion recognition. Huang *et al.*[11] utilized the transformer model to combine acoustic-visual modalities on the model level. First, two separate transformer modules were used to extract features from audio and visual modalities. Then, a multimodal transformer module was adopted to fuse them. Lian *et al.*[12] proposed a multimodal learning framework for conversational emotion recognition, called conversational transformer network. Xie *et al.*[13] investigated a robust approach for multimodal emotion recognition during a conversation. Three separate models for audio, video, and text modalities are structured and fine-tuned. Yu *et al.*[14] utilized a multimodal transformer to improve multimodal named entity recognition.

Inspired by the abovementioned studies, to model the complementarity of linguistic and paralinguistic information, this study proposes an implicitly aligned multimodal transformer fusion (IA-MMTF) framework for emotion recognition based on acoustic and lexical information to promote the learning of multiple complementary emotional representations. The IA-MMTF enables the acoustic and text modalities to guide and complement each other, which adopts a multimodal transformer encoder with cross-modal attention to learn the implicit alignments between the two modalities, thereby learning the acoustic-guided text and text-guided acoustic representations, respectively. Subsequently, the weighted fusion layer is used to further fuse the features of the two modalities to obtain the complementary emotional representation.

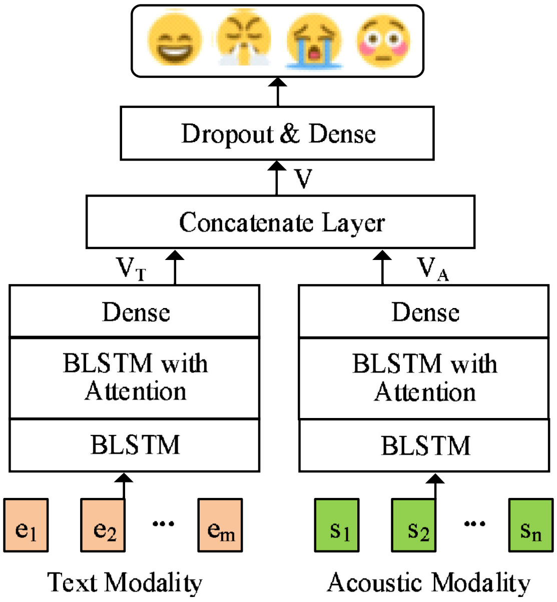


FIGURE 1. Baseline multimodal model: BLSTM with attention.

BASELINE MODEL

Figure 1 shows the baseline multimodal model—BLSTM with an attention mechanism. Compared with LSTM, BLSTM can better capture bidirectional context information. The acoustic modality uses LLD features as inputs, whereas the lexical information is obtained by using Glove embedding. Subsequently, two stacked BLSTM layers followed by a dense layer are adopted to extract acoustic representation V_A and lexical representation V_T from acoustic and text modalities, respectively. In addition, the attention mechanism is integrated with BLSTM to capture important words and time frames. Finally, the concatenate layer is used to fuse the acoustic and text representations

$$V = [V_A, V_T] \tag{1}$$

where V is the fusion representation.

The baseline model is an utterance-level feature fusion framework. Although it integrates the emotional representation of acoustic and text modalities, the lack of fine-grained association study between the two modalities makes it difficult to promote the interaction and influence between different modalities. Thus, the performance of the finally extracted complementary emotional representations will be limited to some extent.

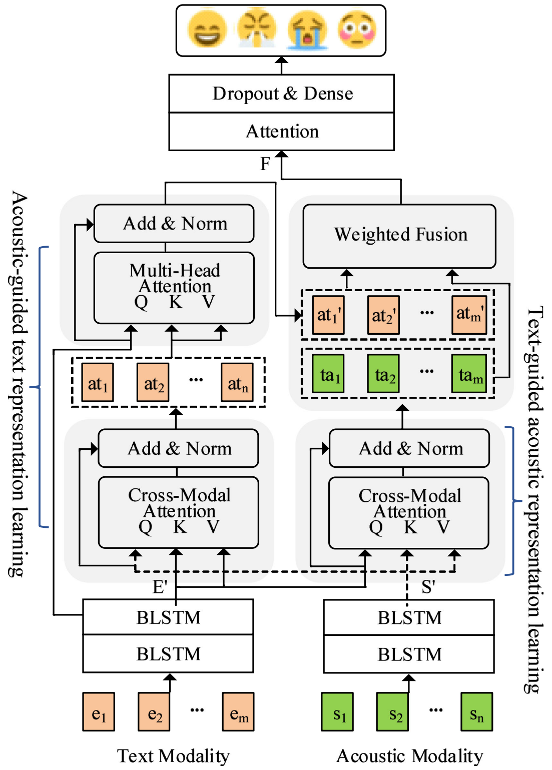


FIGURE 2. Structure of the proposed multimodal transformer fusion framework.

IMPLICITLY ALIGNED MULTIMODAL TRANSFORMER FUSION

We design an IA-MMTF framework, which is integrated with BLSTM, based on acoustic and lexical information. Figure 2 shows the overall structure of the proposed framework, which contains the following three main components:

1) unimodal feature extraction, that is, extraction of acoustic and text features;
2) applying multimodal transformer to implicitly align acoustic and text modalities, to learn the acoustic-guided text representation and text-guided acoustic representation learning based on the multimodal transformer;
3) weighted fusion.

Unimodal Feature Extraction

For the acoustic modality, we use the INTERSPEECH 2010 Paralinguistic Challenge feature set (IS10)[15] to extract 38 LLDs (such as pulse-code modulation loudness and Mel-frequency cepstral coefficients) and their corresponding delta coefficients for a total of 76-dimension. These acoustic features are extracted using the OpenSMILE toolkit.[16] We subsequently zero pad the extra signal and end up with the maximum frame number n for each utterance. Thus, we can obtain the (n,76) input $S = (s_1, s_2, \ldots, s_n)$ for each speech signal, where n denotes the number of time frames, whereas $s_i \in \mathbb{R}^{76}$ denotes the acoustic features for each frame. For the text modality, we adopt the transcript of each speech signal. The pretrained Glove embeddings of dimension 300 along with the maximum sequence length of m is utilized to obtain a (m,300) vector $E = (e_1, e_2, \ldots, e_m)$ for each utterance, where m denotes the number of words and $e_i \in \mathbb{R}^{300}$ is the word vector.

The text input $E \in \mathbb{R}^{m\times 300}$ and acoustic input $S \in \mathbb{R}^{n\times 76}$ are then fed into the two stacked BLSTM layers with d hidden nodes to extract the textual representations $E' = (e'_1, e'_2, \ldots, e'_m)$ and acoustic hidden representations $S' = (s'_1, s'_2, \ldots, s'_n)$, respectively, where $e'_i \in \mathbb{R}^{2\ d}$, $s'_i \in \mathbb{R}^{2\ d}$.

Implicit Alignment

In contrast to the baseline model that extracts the text and acoustic representation independently without any interactions, the proposed model can capture the intermodality dynamics between the lexical and acoustic representations. Two multimodal transformer components are used to learn the implicit alignments between acoustic and text modalities. The multimodal transformer component contains parts of the transformer encoder that include cross-modal attention, residual connection, and layer normalization.

Acoustic-guided text representation: When extracting text representations, we consider the influence of the associated acoustic information. We use acoustic features to guide the text modality to learn the lexical representations with more obvious emotional expressions. First, we adopt h heads cross-modal attention with $S' \in \mathbb{R}^{n\times 2\ d}$ as queries, and $E' \in \mathbb{R}^{m\times 2\ d}$ as keys and values as follows:

$$\mathrm{Att}(S,' E') = \mathrm{softmax}(\frac{(S'W_q)(E'W_k)^T}{\sqrt{2\ d/h}})(E'W_v) \qquad (2)$$

where $W_q, W_k, W_v \in \mathbb{R}^{2\ d\times 2\ d/h}$ are the weights for the query, key, and value. The attention linearly projects the queries, keys, and values h times with different learned linear projections. The h results are concatenated and projected as follows:

$$M_h(S,' E') = \mathrm{Concat}(\mathrm{Att}_1, \mathrm{Att}_2, \ldots, \mathrm{Att}_h)W \qquad (3)$$

where Att_i denotes the ith cross-modal attention, $W \in \mathbb{R}^{2\ d\times 2\ d}$.

We then adopt the residual connection and layer normalization

$$\mathrm{AT} = \mathrm{LayerNorm}(S' + M_h(S,' E')) \tag{4}$$

where $\mathrm{AT} = (at_1, at_2, \ldots, at_n)$ is the acoustic-guided text representation.

However, we can observe that the sequence number of the text representation $\mathrm{AT} \in \mathbb{R}^{n \times 2\ d}$ corresponds to the acoustic modality because the acoustic features are used as queries. To extract the text representations corresponds to each word, we apply another transformer encoder, which adopts E' as queries, and AT as keys and values. As shown in the top left of Figure 2, the new acoustic-guided text representations $\mathrm{AT}' \in \mathbb{R}^{m \times 2\ d}$ are generated.

Text-guided acoustic representation: To obtain the acoustic representation corresponding to each word, thereby reducing the interference of silent segment, we apply a multimodal transformer by treating E' as queries, and S' as keys and values

$$\mathrm{Att}(E,' S') = \mathrm{softmax}(\frac{(E'W_q)(S'W_k)^T}{\sqrt{2\ d/h}})(S'W_v) \tag{5}$$

$$M_h(E,' S') = \mathrm{Concat}(\mathrm{Att}_1, \mathrm{Att}_2, \ldots, \mathrm{Att}_h)W \tag{6}$$

$$\mathrm{TA} = \mathrm{LayerNorm}(E' + M_h(E,' S')) \tag{7}$$

where $\mathrm{TA} = (ta_1, ta_2, \ldots, ta_m)$ is the text-guided acoustic representation.

Weighted Fusion

Paralinguistic information contains different emotional information from linguistic information, and the two modalities may play different roles in emotional expression for each sentence. Therefore, the weights between different modalities need to be considered to obtain richer comprehensive emotional representation. In addition, after the multimodal transformer encoder, the acoustic representations retain the same sequences as the text representations. Thus, we adopt the weighted fusion layer to control the contributions of AT' and TA for each time sequence. The weighted fusion is expressed as follows:

$$F = W_1.\mathrm{AT}' + W_2.\mathrm{AT} \tag{8}$$

where W_1 and W_2 are weight matrices for text and acoustic representations, respectively. We first initialize the weighted matrices randomly and then learn the optimal values during the training. $F \in \mathbb{R}^{m \times 2\ d}$ denotes the final fusion representations for each sequence.

As emotions reflect in different temporal sequences, we then adopt the same attention mechanism with the baseline model to focus on emotion-informing sequences.

EXPERIMENTS AND ANALYSIS

In this section, we first introduce the used database. Then, the experimental setup is described. Finally, the experimental results and analysis are given.

Database

We use the interactive emotional dyadic motion capture (IEMOCAP)[17] and multimodal emotionlines dataset (MELD)[18] databases to validate the proposed method. IEMOCAP and MELD are two publicly available databases for emotion recognition that contain both acoustic and text modalities.

The IEMOCAP database[17] contains approximately 12 h of recordings, including video, speech, and text transcriptions. There are five sessions, and each session was performed by two speakers. In total, 10 American-English speaking people split into five pairs participated in the process. The sentences are annotated in terms of anger, happiness, sadness, neutrality, frustration, excitement, fear, surprise, and disgust by three different human annotators. People usually use the utterances with at least two agreed-upon emotion labels for experiments. Finally, it is common for IEMOCAP to be evaluated as the following four classes: happiness (1636), sadness (1084), anger (1103), and neutrality (1708). We conduct speaker-independent experiments. The first four sessions are adopted as the training set, whereas the remaining session is used as the test set, which is similar to the study.[6]

The MELD database[18] contains 13.7 h of dialogue scenarios from Friends TV series. Different from IEMOCAP that contains dyadic conversations, MELD is a multiparty dataset where multiple speakers participated in the dialogues. There are around 13,708 utterances with seven emotions—anger (1607), disgust (361), sadness (1002), joy (2308), neutral (6436), surprise (1636), and fear (358).

Experimental Setup

The hyperparameters are tuned by grid search and chosen with the best performance. For acoustic modality, the maximum frame number is set to 400 to extract the acoustic features. For text modality, the maximum word number is set to 100. In the multimodal transformer encoder, all the attention layers consist of eight heads. All the BLSTM layers contain 128 hidden nodes. The number of hidden nodes of all dense layers is set to 256. The dropout layer with a 0.25 rate is adopted in our model. When training the model, we select cross-entropy as the cost function and use the Adam optimization algorithm.

To demonstrate the efficiency of the proposed method, we designed multiple groups of comparative experiments, mainly including unimodal emotion recognition methods and the baseline multimodal method: BLSTM with attention. All the experiments list as follows.

- *BLSTM+attention (A)*—This is a unimodal method that uses only acoustic modality. It first extracts the LLD features and then feeds them to two BLSTM layers with 128 nodes to extract the deep representations. Among them, the second BLSTM layer is integrated with self-attention for paying attention to the time frames with obvious emotional features.
- *BLSTM+attention (T)*—This method uses only text modality, which first adopts the pretrained Glove embeddings to obtain word vectors. Thereafter, the same BLSTM layers as the acoustic modality are used to learn text representations
- *BLSTM+attention (A+T)*—This is the baseline model that uses both acoustic and text modalities for emotion recognition. The structure is shown in Figure 1.
- *BLSTM+IA-MMT (A+T)*—This is an ablation study that uses only the implicitly aligned multimodal transformer (IA-MMT) without using weighted fusion. It adopts the fully connected layer to combine audio and textual representations.
- *BLSTM+IA-MMTF (A+T)*—This is the proposed multimodal framework. First, the acoustic and text modalities are implicitly aligned. Thereafter, the weighed fusion layer is adopted to further fuse them.

Validation of the Proposed Framework

Table 1 lists the comparison results on the IEMOCAP database using two common assessment criteria, weighted accuracy (WA) and unweighted accuracy (UA). WA measures the classification accuracy of all test utterances. UA is the average classification accuracy for each emotion. From Table 1, a summary of conclusions can be drawn as follows.

- Compared with the unimodal methods, the baseline model "BLSTM+attention" exhibits notable improvements. It showed absolute improvements of 3.63% (from 65.83% to 69.46%) and 3.38% (from 67.15% to 70.53%) in terms of WA and UA, respectively. The results indicate that there are clear complementarities between acoustic and text modalities.
- The results also indicate that the ablation study "BLSTM+IA-MMT" outperforms the baseline model, which suggests that using the multimodal transformer for implicit modal alignment can improve emotion recognition. But the results are lower than that of "BLSTM+IA-MMTF," which also verifies the effect of weighted fusion.
- The proposed "BLSTM+IA-MMTF" is highly competitive to the baseline multimodal model. It significantly outperforms the baseline model by absolute improvements of 2.50% (from 69.46% to 71.96%) and 1.96% (from 70.53% to 72.49%) in terms of WA and UA, which indicates that the proposed method is more advantageous in the fusion of acoustic and text modalities. The reason may be that our model not only learns the mutual relationships between different modalities but also uses weighted fusion to effectively control the contributions of each modality.

TABLE 1. Comparison with baseline results on the IEMOCAP database: "a" and "t" indicate "acoustic" and "text."

Method	**Modality**	**WA(%)**	**UA(%)**
BLSTM+attention	A	56.00	56.40
BLSTM+attention	T	65.83	67.15
BLSTM+attention	A + T	69.46	70.53
BLSTM+MMT	A + T	70.91	71.84
BLSTM+IA-MMTF	A + T	**71.96**	**72.49**

The bold values indicate the best performance of each column.

To analyze the contribution of our methods on classifying different types of emotion, Figure 3 gives F1 scores of four emotions. F1 is the most commonly used evaluation criterion for testing accuracy because it can keep the balance between recall (R) and precision (P).

From Figure 3, we can observe that the proposed method "BLSTM+IA-MMTF" performs best in most classes including neutrality, anger, and happiness, and shows the similar performance to the baseline model in sadness. For average F1, the proposed method outperforms the acoustic modality, text modality, and baseline multimodal by absolute improvements of 16.84%, 6.48%, and 2.69%, respectively. These results demonstrate the effectiveness of the proposed multimodal transformer fusion framework, and the learned emotional representations can better utilize the complementarity between acoustic and text modalities.

To further analyze the effects of unimodal and multimodal methods on each emotion, Figure 4 illustrates the confusion matrices of the "BLSTM+attention (A)," "BLSTM+attention (T)," baseline model "BLSTM+attention (A+T)," and the proposed method "BLSTM+IA-MMTF (A+T)" on the IEMOCAP database. In the figures,

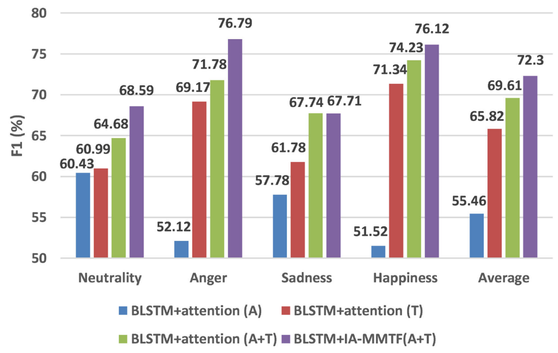
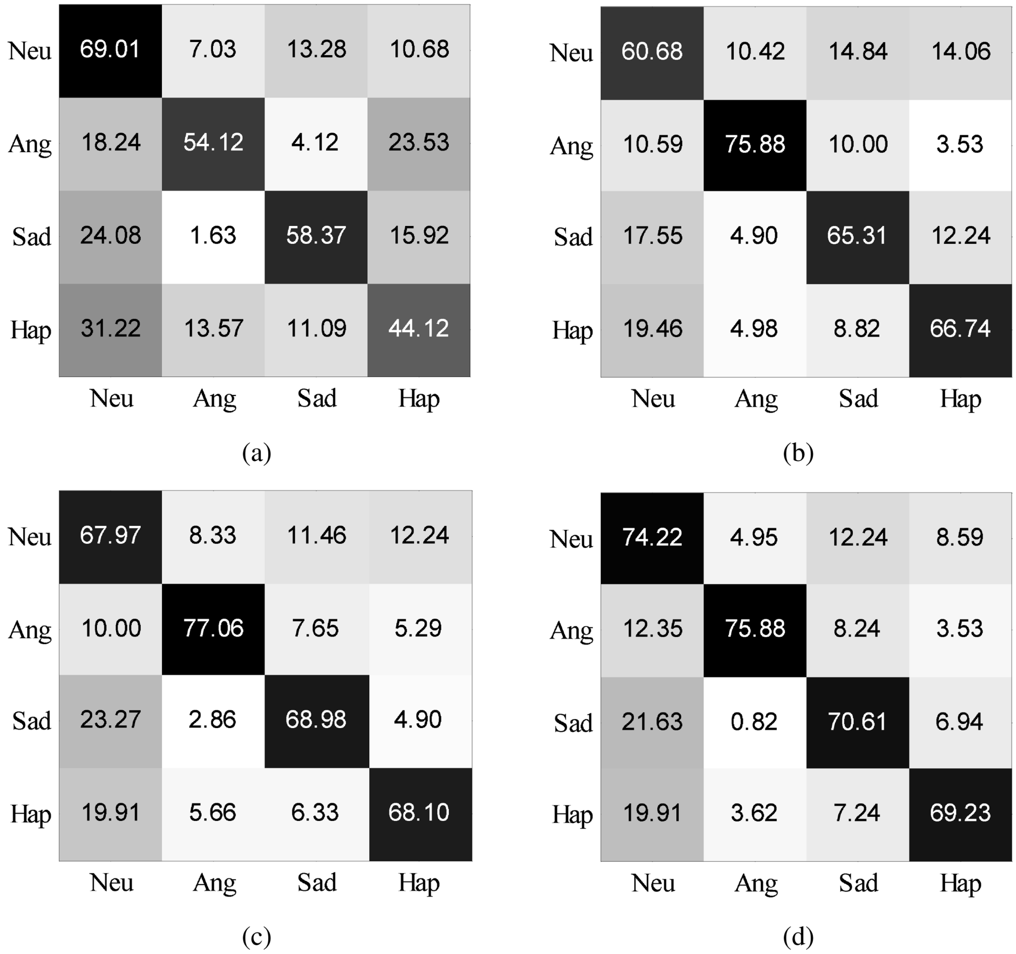


FIGURE 3. F1(%) on the IEMOCAP database.

the abscissas show the predicted values, whereas the ordinates show the actual values.

We can observe that text modality [see Figure 4(b)] performs better than acoustic modality [see Figure 4(a)] in classifying anger, sadness, and happiness, especially for happiness (from 44.12% to 66.74%) and anger (from 54.12% to 75.88%), the absolute improvements are more than 20%. But for neutrality, the acoustic modality outperforms the text modality by 8.33% (from 60.69% to 69.01%) improvements. The main reason is probably that neutrality is relatively implicit in linguistic characteristics. In addition, noticeable confusion occurs between happiness and anger in Figure 4(a), but not in Figure 4(b). That is probably because people usually express happiness and anger by using words with clear emotional tendencies; thus, their lexical representations are relatively easy to distinguish. However, there are some commonalities for the acoustic modality; for instance, both happiness and anger have high arousal and energy.

For multimodal methods, the baseline model improves most emotions, but the accuracy of neutrality is worse than that of the acoustic modality, as shown in Figure 4(c). This indicates that this method



	Neu	Ang	Sad	Hap
Neu	69.01	7.03	13.28	10.68
Ang	18.24	54.12	4.12	23.53
Sad	24.08	1.63	58.37	15.92
Hap	31.22	13.57	11.09	44.12

(a)

	Neu	Ang	Sad	Hap
Neu	60.68	10.42	14.84	14.06
Ang	10.59	75.88	10.00	3.53
Sad	17.55	4.90	65.31	12.24
Hap	19.46	4.98	8.82	66.74

(b)

	Neu	Ang	Sad	Hap
Neu	67.97	8.33	11.46	12.24
Ang	10.00	77.06	7.65	5.29
Sad	23.27	2.86	68.98	4.90
Hap	19.91	5.66	6.33	68.10

(c)

	Neu	Ang	Sad	Hap
Neu	74.22	4.95	12.24	8.59
Ang	12.35	75.88	8.24	3.53
Sad	21.63	0.82	70.61	6.94
Hap	19.91	3.62	7.24	69.23

(d)

FIGURE 4. (a) Confusion matrix of acoustic modality on IEMOCAP; (b) Confusion matrix of text modality; (c) Confusion matrix of the baseline method; (d) Confusion matrix of the proposed method.

TABLE 2. Comparison with baseline results on the MELD database.

Method	Modality	WA(%)	WF1(%)
BLSTM+attention	A	48.70	38.20
BLSTM+attention	T	52.72	45.99
BLSTM+attention	A + T	53.56	47.08
BLSTM+MMT	A + T	54.18	47.93
BLSTM+IA-MMTF	A + T	**54.79**	**48.96**

The bold values indicate the best performance of each column.

has limitations in utilizing the complementarity between acoustic features and text information, and cannot make full use of paralinguistic information to supplement linguistic information. However, compared with unimodal methods, our method performs best on all emotions in Figure 4(d), especially, for neutrality and sadness with absolute improvements of over 5%. Furthermore, our method outperforms the baseline model in neutrality, sadness, and happiness, especially for neutrality with improvements of 6.25% (from 67.97% to 74.22%). This shows that the proposed method can effectively utilize the complementarity of the two kinds of information: when it is difficult to express emotion with linguistic information, it can be supplemented by paralinguistic information; meanwhile, the emotions which are difficult to be expressed by paralinguistic information can be realized by linguistic information.

Table 2 shows the results on the MELD database using the following two metrics: WA and weighted average F1 (WF1). WF1 is a weighted mean F1 over different emotion categories with weights proportional to the number of utterances in a particular emotion class, which is suitable for the MELD database with exceeding unbalanced data distribution.

From Table 2, we can observe that our method outperforms the baseline multimodal method by absolute improvements of 1.23% (from 53.56% to 54.79%) and 1.88% (from 47.08% to 48.96%) in terms of WA and WF1, which demonstrates that the proposed method is effective on the MELD database.

Figure 5 gives F1 scores of each emotion on the MELD database. The proportion of "disgust" and "fear" is lower than 3%, resulting in results of 0%. Thus, we do not report their F1 scores in Figure 5. From this figure, we can observe that the proposed method "BLSTM+IA-MMTF" performs best in most classes including neutra, anger, and joy, and shows the second-highest result in surprise emotion.

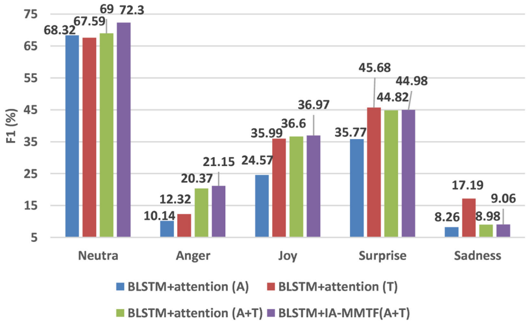


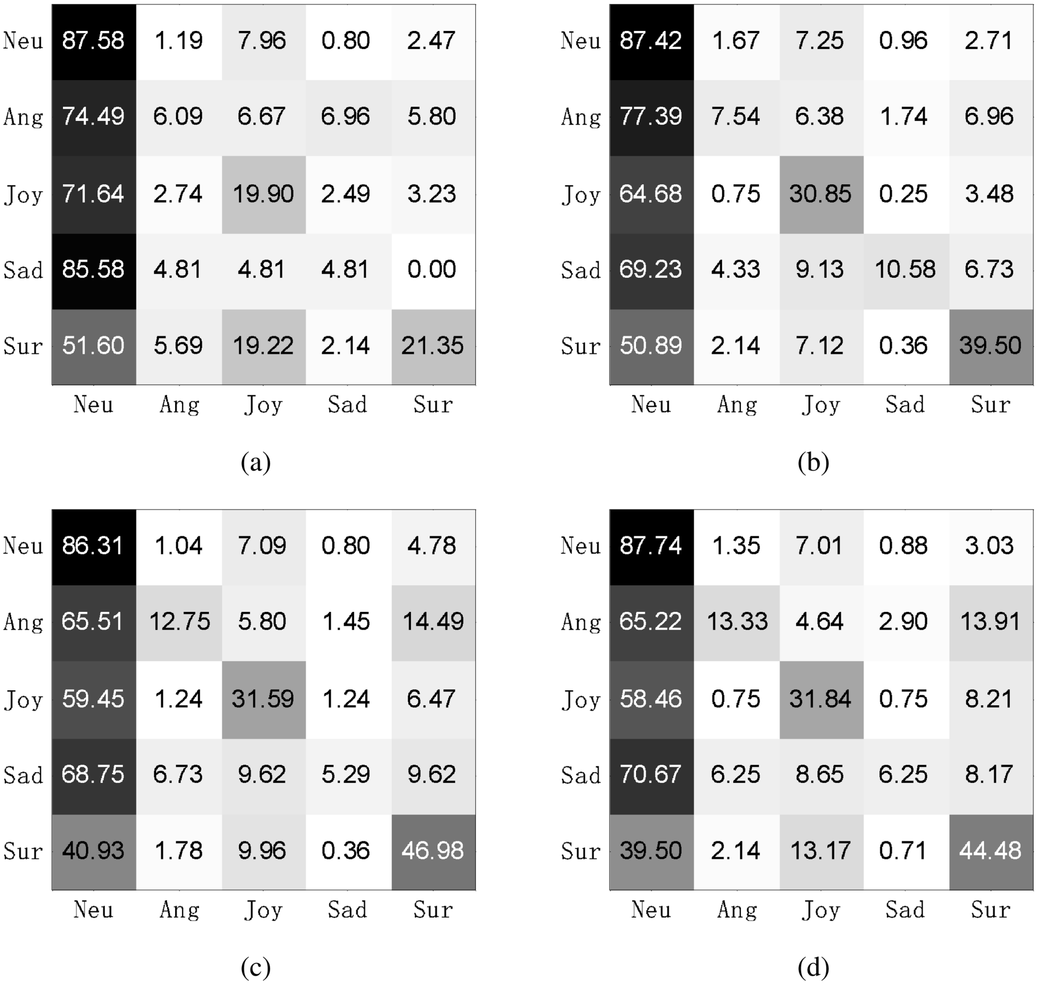
FIGURE 5. F1(%) on the MELD database.

We also illustrate the confusion matrices in Figure 6 to further analyze the effects of unimodal and multimodal methods on each emotion. We can observe that text modality [see Figure 6(b)] performs better than acoustic modality [see Figure 6(a)] in most classes except neutral emotion, which is similar to IEMOCAP database. We can also observe that all the utterances are easy detected as neutral class beacuse MELD is a highly unbalanced database and "nuetral" has the highest proportion. Compared with unimodal methods, our method performs best on most emotions in Figure 4(d). Furthermore, our method outperforms the baseline model in neutral, anger, joy, and sad.

Comparison With Other Methods

There are many multimodal methods based on acoustic and lexical information for emotion recognition. In the study of Jin *et al.*,'s work,[4] a new representation from Gaussian Supervectors is extracted in acoustic modality and a new feature representation named emotion vector (eVector) is in text modality, thereby fusing them in decision level. Cho *et al.*[9] proposed a late fusion network using LSTM MCNN.

Gamage *et al.*[5] combined acoustic features LLD with the modified relative frequency-based lexical feature for multimodal emotion recognition. Sahu *et al.*[19] proposed two integration methods based on acoustic and lexical information, including E1–Ensemble (Random Forest + Gradient Boosting + Multilayer Perceptron) and E2–Ensemble (Random Forest + Gradient Boosting + Multilayer Perceptron + Multinomial Naive Bayes + Logistic Regression). "BLSTM+attention" is the baseline model that referred to previous studies.[6,7] Considering that most previous methods on the MELD database were proposed for conversational emotion recognition, there is less comparability between our results and these



	Neu	Ang	Joy	Sad	Sur
Neu	87.58	1.19	7.96	0.80	2.47
Ang	74.49	6.09	6.67	6.96	5.80
Joy	71.64	2.74	19.90	2.49	3.23
Sad	85.58	4.81	4.81	4.81	0.00
Sur	51.60	5.69	19.22	2.14	21.35

(a)

	Neu	Ang	Joy	Sad	Sur
Neu	87.42	1.67	7.25	0.96	2.71
Ang	77.39	7.54	6.38	1.74	6.96
Joy	64.68	0.75	30.85	0.25	3.48
Sad	69.23	4.33	9.13	10.58	6.73
Sur	50.89	2.14	7.12	0.36	39.50

(b)

	Neu	Ang	Joy	Sad	Sur
Neu	86.31	1.04	7.09	0.80	4.78
Ang	65.51	12.75	5.80	1.45	14.49
Joy	59.45	1.24	31.59	1.24	6.47
Sad	68.75	6.73	9.62	5.29	9.62
Sur	40.93	1.78	9.96	0.36	46.98

(c)

	Neu	Ang	Joy	Sad	Sur
Neu	87.74	1.35	7.01	0.88	3.03
Ang	65.22	13.33	4.64	2.90	13.91
Joy	58.46	0.75	31.84	0.75	8.21
Sad	70.67	6.25	8.65	6.25	8.17
Sur	39.50	2.14	13.17	0.71	44.48

(d)

FIGURE 6. (a) Confusion matrix of acoustic modality on MELD; (b) Confusion matrix of text modality; (c) Confusion matrix of the baseline method; (d) Confusion matrix of the proposed method.

studies. Thus, the methods that can direct comparisons with us are limited on the MELD database.

Tables 3 and 4 list the comparison results on the IEMOCAP and MELD database, respectively. We can observe that the results of the proposed method are better than that of other methods, which indicates that our method can effectively use the complementarity of acoustic and lexical information.

CONCLUSION

Considering that emotion in speech is expressed jointly by paralinguistic and linguistic information, this article modeled the complementarity of them, and proposed a multimodal transformer fusion (IA-MMTF) framework across acoustic-text modality. The transformer with cross-modal attention was adopted to perform implicit alignments between acoustic and text modalities, thereby enabling the two modalities to guide and complement each other. Thereafter, weighted fusion was utilized to further

TABLE 3. Comparison results of our method and other methods on the IEMOCAP database.

Method	WA(%)	UA(%)
LLD+BoW+GSV_{mean}+eVector[4]	69.20	68.65
eVector+MCNN+LSTM[9]	64.90	65.90
E1:(RF+XGB+MLP)[19]	70.30	65.50
E2:(RF+XGB+MLP+MNB+LR)[19]	70.10	71.50
BLSTM + attention[6,7]	69.46	70.53
BLSTM + IA-MMTF	**71.96**	**72.49**

The bold values indicate the best performance of each column.

TABLE 4. Comparison results of our method and other methods on the MELD database.

Method	WA(%)	WF1(%)
LLD+mLRF[5]	50.19	44.03
CNN[20]	52.15	45.45
BLSTM + attention[6,7]	53.56	47.08
BLSTM + IA-MMTF	**54.79**	**48.96**

The bold values indicate the best performance of each column.

integration for learning more complementary emotional representations. The results on the IEMOCAP and MELD databases demonstrate that the proposed method significantly outperforms the baseline multimodal model, which verifies that our method can integrate acoustic and lexical information efficiently. Further study on the details of complementation will be our future work.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grants 62176182 and 61976216, and in part by the Tianjin Municipal Science and Technology Project under Grant 19ZXZNGX00030.

REFERENCES

1. H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users' affective states," *Appl. Artif. Intell.*, vol. 19, no. 3/4, pp. 267–285, Mar. 2005.
2. E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6720–6724.
3. J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proc. 7th ACM Annu. Workshop*, 2017, pp. 11–18.
4. Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4749–4753.
5. K. W. Gamage, V. Sethu, and E. Ambikairajah, "Salience based lexical features for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5830–5834.
6. S. Tripathi, S. Tripathi, and H. Beigi, "Multimodal emotion recognition on IEMOCAP dataset using deep learning," 2019, *arXiv:1804.05788v3*.
7. C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, pp. 1–9, May 2020.
8. Y. Gu, S. Chen, and I. Marsic, "Deep multimodal learning for emotion recognition in spoken language," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5079–5073.
9. J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. INTERSPEECH*, 2018, pp. 247–251.
10. A. Vaswani *et al.*, "Attention is your need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
11. J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3507–3511.
12. Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, 2021.
13. B. Xie, B. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, Jul. 2021.
14. J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3342–3352.
15. B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.
16. F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
17. C. Busso, M. Bulut, and C. H. Lee, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
18. S. Poria *et al.*, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 527–536.
19. G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," 2019, *arXiv:1904.06022*.

20. Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

LILI GUO is a lecturer at the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China, and also with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China. Her research interests include emotion recognition, acoustic signal processing, and deep learning. Guo received her Ph.D. degree from Tianjin University. Contact her at liliguo@cumt.edu.cn.

LONGBIAO WANG is a professor at Tianjin University, Tianjin, 300350, China. From 2008 to 2012, he was an assistant professor with Shizuoka University, Shizuoka, Japan. From 2012 to 2016, he was an associate professor with the Nagaoka University of Technology, Nagaoka, Japan. His research interests include speech recognition and acoustic signal processing. He is a member of IEEE. He is the corresponding author of this article. Contact him at longbiao_wang@tju.edu.cn.

JIANWU DANG is a professor with Tianjin University, Tianjin, 300350, China, and also at Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, 923-1211, Japan. Since 2001, he has been a professor with the Faculty of JAIST. His research interests include speech production/synthesis/recognition. Dang received his Ph.D. degree from Shizuoka, Japan. He is a member of IEEE. Contact him at jdang@jaist.ac.jp.

YAHUI FU is a researcher at the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 923-1211, Japan. Her research interests include multimodal emotion recognition and spoken dialogue system. Fu received her M.S. degree from both Tianjin University, Tianjin, China, and Japan Advanced Institute of Science and Technology. Contact her at s1910275@jaist.ac.jp.

JIAXING LIU is currently working toward the Ph.D. degree with the College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China. His research interests include speech emotion recognition and multimodal emotion recognition. Contact him at jiaxingliu@tju.edu.cn.

SHIFEI DING is a professor at the China University of Mining and Technology, Xuzhou, 221116, China. He was a postdoctoral fellow of the Chinese Academy of Sciences, Beijing, China. His research interests include artificial intelligence and intelligent information processing. He is the corresponding author of this article. Contact him at dingsf@cumt.edu.cn.