INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Dual Attention-based Modality-Collaborative Fusion Network for Emotion Recognition

*Xiaoheng Zhang, Yang Li**

Beihang University, Beijing, China
{xiaoheng_zhang, liyang}@buaa.edu.cn

Abstract

Multi-modal emotion recognition (MER) is an emerging research field in human-computer interactions. However, previous studies have explored several fusion methods to deal with the asynchronism and the heterogeneity of multimodal data but they mostly neglect the importance of discriminative unimodal information resulting in the ignorance of independence of uni-modality. Furthermore, the complementarity among different fusion strategies is seldom taken in consideration. To address these limitations, we propose a modality-collaborative fusion network (MCFN) consisting of three main components: a dual attention-based intra-modal learning module which is devoted to build the initial embedding spaces, a modality-collaborative learning approach is to reconcile the emotional information across modalities, and a two-stage fusion strategy to integrate multimodal features which are improved by a mutual adjustment approach. The proposed framework outperforms the state-of-the-art methods in overall experiments on two well-known public datasets. Our model will be available at https://github.com/zxiaohen/Speech-emotion-recognition-MCFN

Index Terms: Multimodal emotion recognition, Intra-modal, Modality-collaborative

1. Introduction

Emotion plays a crucial role in our daily communication and emotion recognition finds applications in different domains like customer services [1], social robots and dialogue systems. In recent years, multimodal emotion recognition (MER) has attracted increasing attention. Firstly, various deep learning approaches have been applied to improve the performance of speech emotion recognition. Deep learning approaches such as bidirectional LSTMs in combination with attention mechanism [2], and time-delay neural networks (TDNN) [3] are applied in order to capture contextual information. In addition, a lightweight 1D CNN SER system [4], which utilizes the dilated convolution layers, shows its performance in capturing the local low-level features as the computational speed increases. However, when the emotion expressed through speech becomes ambiguous, the lexical content may provide complementary information that can address the ambiguity. As for text encoder approaches, Word2Vec and Glove [5] are two widely used unsupervised word embeddings, however, they have limitations in capturing contextual information, our study uses a finetuned RoBERTa model [6] for text emotion recognition which are more capable to capture the context with the position embeddings and multi-layer transformers. Furthermore, one of the challenges is to determine the most relevant acoustic and semantic emotion features. Therefore, a temporal and channel-wise dual attention structure may be more appropriate to address this issue.

The early research focused on modality-independent fusion which combined the acoustic and semantic features directly. Many researchers [7, 8] have also explored modality-independent late fusion which is easier to handle. However, the correlation between features and emotionally-salient words was rarely considered. To solve these drawbacks, the modality-dependent approaches were investigated, many researchers have introduced the attention mechanism between words and frames to achieve an alignment considering the utterance-level global influence [9], and various cross-modal interaction structures have been proposed [10, 11] with an attention module to learn representations from unaligned multimodal sources. However, most of the existing works only focus on either feature fusion which can effectively exploit the covariations between features from different modalities, or decision fusion, which shows the robustness of capturing an optimal combination of two modalities and errors from multiple models are dealt with independently. This observation motivates us to design a hybrid fusion method. In this paper, we propose a Modality-Collaborative Fusion Network (MCFN), the main contributions can be summarized as follows:

- We introduce an Intra-modal Learning Module to ensure the independence of the single modality and at the extraction of the emotion-related features in both the channel and temporal directions by the use of a dual attention mechanism.
- We design a Modality Collaborative Learning Module to obtain semantic information from other modalities considering the cross-modal alignment and interaction.
- We introduce a two-stage fusion approach with a well-designed mutual adjustment to merge progressively emotional information from intra-/inter-modality.

2. Proposed Multimodal Framework

Fig.1 shows the block diagram of the proposed two-stage fusion network including intra-modal and modality-collaborative learning. Our model consists of three components: (a) ILM (Intra-modal Learning Module) where a dual attention block is integrated, (b) MCLM (Modality-Collaborative Learning Module) including a multi-head co-attention alignment and a cross-modal interaction, followed with a mutual adjustment and (c) the emotion classifier where we combine the information of both modalities effectively and get final predictions by a decision-level fusion. The following subsections provide a description of the structures in this framework.

*Corresponding author

10.21437/Interspeech.2023-523

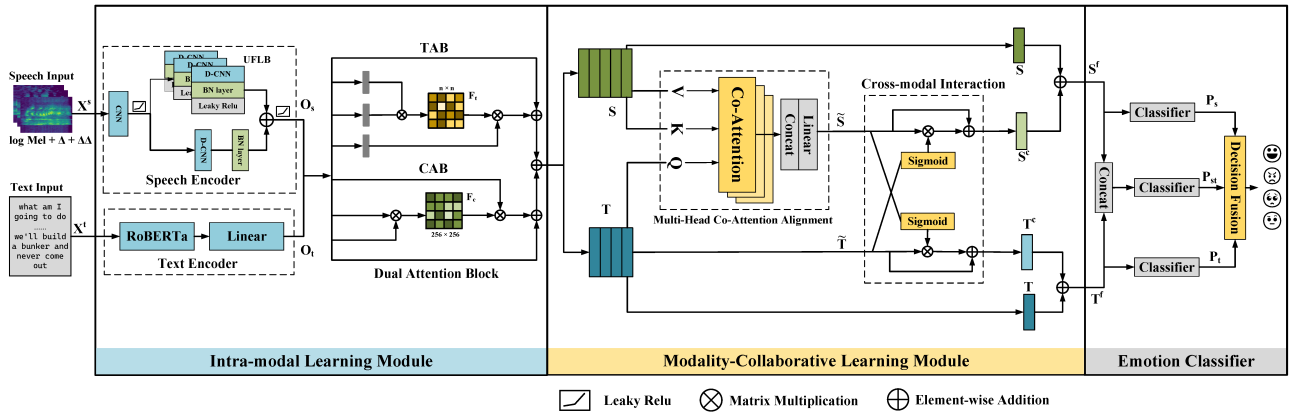


Figure 1: *Overall architecture of proposed MCFN. We compute intra-modal learning (ILM) and modality-collaborative learning (MCLM) for two modalities, and merge the cross-modal features with original unimodal features in a hierarchical way. TAB represents temporal-wise attention branch while CAB represents channel-wise attention branch.*

2.1. Intra-modal Learning Module

In this work, we utilize the frame-level and word-level features as input. The raw acoustic input is represented as X^s while the semantic input is X^t. We leverage the spectrogram augmentation technique by applying random time-frequency masks to spectrograms as described in [12]. Then, considering dynamic information about transitions between frames, we use the first and second derivatives of log-MFB features, which are called delta and delta-delta coefficients [13]. Then a ligntweight 1D dilated convolutional network is performed including three upgrade feature learning blocks (UFLB) as illustrated in Fig.1, a combination of one dilated convolutional layer, one batch normalization (BN) layer and one Leaky-ReLU layer, and a skip connection residual block is applied to integrate the current information with the previous one. We extract the linguistic representations via pre-trained and finetuned RoBERTa model followed by a fully-connected layer. Finally the acoustic embedding from intra-modal learning is represented as $O_s \in \mathbb{R}^{q\times n}$, $O_s = \{O_{s_1}, O_{s_2}...O_{s_n}\}$ while the textual embedding can be represented as $O_t \in \mathbb{R}^{d\times m}$, $O_t = \{O_{t_1}, O_{t_2}...O_{t_m}\}$ where n, m denotes the length of the sequence, $q, d = 256$ represent the dimensionality of embedding spaces.

For each modality, we learn high-quality latent representations by encoding followed by a temporal and channel-wise dual attention to capture features dependencies in the temporal and channel dimensions respectively. As illustrated in Fig.1, we design two types of attention modules to draw global context over local features, thus obtaining better feature representations for word-level prediction. Two branches are designed as TAB (Temporal-wise Attention Branch) which upgrades the feature along time-series and encodes the contextual information into local features in form of a weighted sum, and CAB (Channel-wise Attention Branch) which emphasizes interdependent channel maps by differentiate the importance of different channels. O_s and O_t go into the dual attention independently and we obtain respectively S and T.

As shown in Fig.1, firstly we feed the feature embedding O_t into convolution layers to generate two feature maps T_1, T_2 respectively and reshape them, then we perform a matrix multiplication and a softmax layer to calculate the temporal attention map $F_t = (t_{i,j})_{m\times m} \in \mathbb{R}^{m\times m}$. We feed O_t into a convolution layer to generate a new feature map T_3 and perform a matrix multiplication with F_t (Eq.(1)). Finally, we perform an element-wise sum operation to obtain the final output. Similar to TAB, we build a channel attention branch by directly calculating the channel attention map $F_c = (c_{i,j})_{256\times 256} \in \mathbb{R}^{256\times 256}$ (Eq.(2)), and we aggregate the features from TAB and CAB to obtain $T = T_t + T_c \in \mathbb{R}^{d\times m}$ and $S = S_t + S_c \in \mathbb{R}^{q\times n}$.

$$t_{j,i} = \frac{exp(T_{1i} \cdot T_{2j})}{\sum_{t=1}^{N} exp(T_{1i} \cdot T_{2j})}, \quad T_t = T + \sum_{i=1}^{m} t_{i,j} T_3 \tag{1}$$

$$c_{j,i} = \frac{exp(T_i \cdot T_j)}{\sum_{t=1}^{N} exp(T_i \cdot T_j)}, \quad T_c = T + \sum_{i=1}^{256} c_{i,j} T \tag{2}$$

where $t_{j,i}$ measures the i^{th} temporal state's impact on the j^{th} and $c_{i,j}$ measures the i^{th} channel's imapact on the j^{th} channel.

2.2. Modality-Collaborative Learning Module

To increase more discrimination in emotional cues across modalities, the alignment between speech frames and text words after the dual attention is completely learned from the co-attention mechanism. Given an encoded speech embedding S and text embedding T, the high-level acoustic feature can be learned by a multi-head co-attention mechanism where the alignment score between the i^{th} speech frame and the j^{th} word is calculated as Eq.(3):

$$a_{i,j} = tanh(U^T s_i + V^T t_j + b) \tag{3}$$

where U, V and b are trainable parameters. Here, We use the text T as the query because text modality has better robustness because of its rich semantic information compared with audio modality and is less likely to be disturbed by noise interference.

Then we can obtain the normalized attention weight over the speech sequence $\alpha_{i,j}$ indicating the alignment score between the i^{th} frame and the j^{th} word, and the aligned speech feature $\widetilde{S} = \{\widetilde{s}_j\}_{j=1}^{n}$ is projected into the latent space grounded on the text features, in order to facilitate the interpretability of symbols, the text feature after alignment is denoted as $\widetilde{T} = T$:

$$\alpha_{i,j} = \frac{exp(a_{j,i})}{\sum_{t=1}^{N} exp(a_{j,t})}, \quad \widetilde{s}_j = \sum_{i} \alpha_{i,j} s_i \tag{4}$$

Besides, we design a cross-modality interaction block to learn the interaction between two modalities. We obtain two excitement weight matrices of corresponding features. Then, the

aligned embeddings are calibrated to obtain S^c and T^c.

$$S^c = \widetilde{S} + \widetilde{S} \odot E_s, \quad E_s = \sigma(W_s \cdot \widetilde{T}) \tag{5}$$

$$T^c = \widetilde{T} + \widetilde{T} \odot E_t, \quad E_t = \sigma(W_t \cdot \widetilde{S}) \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and W_s, W_t represent linear projection operators.

As for the feature-level local fusion, to ensure that the MCLM can learn from the uni-modality without losing too much original information, instead of directly concatenating the features of the two modalities, we introduce a mutual adjustment specifically a mean squared error function to constrain the difference between intra-modal representations (S, T) and inter-modal representations (S^c, T^c), as shown in Eq.(7). Afterwards, to consider more original information and achieve a non-redundancy, we feed them into a adjustment function which can be of any form that keeps the shape of representation unchanged. We adopt a simple addition and finally obtain two ajusted representations S^f and T^f.

$$\mathcal{L}_{ad} = \frac{1}{N}\sum_{n}^{N}(S - S^c)^2 + \frac{1}{N}\sum_{n}^{N}(T - T^c)^2 \tag{7}$$

where N is the number of samples.

2.3. Emotion classifier

In our work, besides of the feature-level fusion, a decision-level fusion is applied to complementarily utilize the reliability measures of speech and text information. The final prediction is calculated as the summation of probabilities from two adjusted intra-modal classifiers and a bimodal classifier:

$$P(y_i|x) = \alpha P(y_i^s|x^s) + \beta P(y_i^t|x^t) + \gamma P(y_i^{st}|x^{st}) \tag{8}$$

where P represents the probability with x representing the given sequences, y_i as the i-th emotional label, and α, β, γ are hyperparameters which satisfy the constraints:

$$\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1 \tag{9}$$

We adopt a cross-entropy loss as classification loss (Eq.(10)) for each classifier to optimize the latent representations.

$$\mathcal{L}_{cl}^m = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_i \cdot logP(y_i^m|x^m), m \in \{s, t, st\} \tag{10}$$

where C is the total number of classes, y_i is annotated emotional label. The total loss $\mathcal{L}$ can be expressed as Eq.(11):

$$\mathcal{L} = \sum_{m\in\{s,t,st\}} \lambda_m \mathcal{L}_{cl}^m + \omega \mathcal{L}_{ad} \tag{11}$$

where λ is the weighting factor and ω is the hyperparameter to balance the two losses. The joint loss function ensures that the model focuses on various modalities and shows its robustness in learning the joint representation across modalities

3. Experiments

3.1. Datasets

In this section, we conduct several experiments to evaluate the effectiveness of our proposed method and compare it with state-of-the-art baselines on two benchmark datasets:

- **IEMOCAP** [14] dataset contains approximately 12 hours of dyadic emotional improvised and scripted conversations (10039 utterances). The labelling of each utterance was determined by 3 annotators as the following categorical labels: anger, happiness, sadness, neutral,excitement, frustration, fear, surprise. We consider the first four labels, in which the excitement category is merged into happiness and the 5-fold leave-one-session-out (LOSO) strategy is adopted.
- **MELD** [15] is a large-scale multi-party conversational dataset which contains more than 13,000 utterances and each utterance is annotated with one of the following labels: anger, joy, sadness, neutral, disgust, fear and surprise.

3.2. Implementation Details and Metric

Our implementation is based on the PyTorch framework. The raw audio signal is transformed into short frames with 25 ms window width and 10 ms frameshift. The dilation rate is set to 2. The pre-trained RoBERTa model is used to encode text and obtain 768-dimensional features. The network is trained using Adam optimizer with a batch size of 32 and a learning rate of 0.0001, the hidden units for multihead attention is set to 128 and the number of head is set to 4. In our study,we have applied the grid search method for the hyperparameter finetuning, we try every combination of values and calculate the performance metrics using a 5-fold cross-validation scheme. The point of the grid that maximizes the average value in cross-validation is the optimal choice.

Since the commonly-used metrics on the two databases are different, we select the universal metrics on each dataset for measurement to conduct a comparison with the literature, most of the previous works have chosen WA and UA for IEMOCAP considering the test sets are slightly imbalanced between different emotion categories while Acc and F1 for MELD. We utilize the commonly used metrics on each dataset: weighted accuracy (WA) and unweighted accuracy (UA) for the IEMOCAP dataset, accuracy (Acc) and weighted F1-score (WF1) for the MELD dataset. As for the baselines, we have selected the state-of-the-art baseline approaches of the same settings respectively for the two datasets to compare with the latest works.

3.3. Results and analysis

3.3.1. Overall Classification Performance

We compare our proposed method with some state-of-the-art baselines (Table 1). It can be observed that our proposed method achieves more excellent performance than the other state-of-the-art methods. On the IEMOCAP dataset, the proposed MCFN obtains WA and UA of 76.01% and 77.84% which are the best scores among the methods listed. Compared with TDNN[3], which fuse audio features with the corresponding text at each frame directly using a self-attention without cross-modal interaction, we focus on the interaction between two modalities by a cross-attention and design a late fusion to make up for the inadequacy of the common-used early fusion combining early and late fusion. In addition, our proposed MCFN outperforms FAF[16] by 3.3% on WA and 5.1% on UA probably due to the fact that FAF adopted a word-level alignment fusion ignoring cross-modal interaction, and is 1.7%, 2.5% higher than the current best methods TSIN[10], KS-TRM[17] which focus on the cross-modal interaction while many useless relationships are generated and disturb the performance of the classifier.

To further prove the performance of MCFN, we then evaluate it on the MELD dataset (Table 2). DialogueRNN[18] and the semi-supervised method[19] lack cross-modal interaction, while CTNet[20] is a transformer-based model and performs better than the above method without interaction, it's also a typical example of most existing works which pay attention to cross-modal interaction and fusion but neglect the independence

of uni-modal features leading to the generation of useless relationships. More importantly, these interactions often destroy the features obtained from intra-modal learning, resulting in the intra-modal characteristics being assimilated by other modalities. To address this issue, we propose the well-designed dual attention for uni-modality and combine the original unimodal features and cross-modal features as the final input for the classifier. Specifically, our proposed framework utilizes the Modality Collaborative Learning to enhance the recognition performance with the Intra-modal Learning and a two-stage fusion to make full use of information from different levels and different scales without losing the original information. We can notice that MCFN outperforms the state-of-the-art baselines although they mainly focus on the utterance-level features in a contextual conversation while we utilize the word-level embeddings.

Table 1: *Model performance comparison on the IEMOCAP dataset. The results of 5-fold cross-validation are presented below. "S" : Speech, "T" : Text, "V" : Video.*

Methods	Modality	WA (%)	UA(%)
FAF (2018) [16]	S+T	72.7	72.7
TSIN (2021) [10]	S+T	74.9	76.6
TDNN (2021) [3]	S+T	75.5	76.6
KS-TRM (2022) [17]	S+T	74.3	75.3
Proposed	S+T	**76.0**	**77.8**

Table 2: *Model performance comparison on the MELD.*

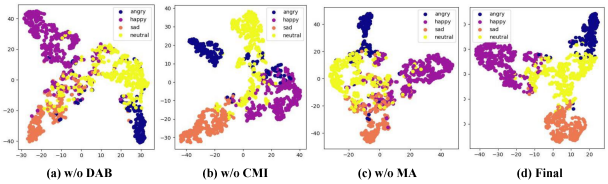
Methods	Modality	Acc (%)	WF1(%)
DialogueRNN (2019) [18]	S+T+V	-	60.3
SSM(2020) [19]	S+T+V	-	57.1
CTNet (2021) [20]	S+T	60.8	62.0
Proposed	S+T	**64.5**	**62.2**

3.3.2. Ablation Studies

In order to explore the contribution of each block, we perform a series of experiments as listed in Table 3. Firstly, we evaluate the effect of modalities and we observe a significant performance drop when we utilize unimodal information as input, and the semantic features outperform acoustic features significantly, which verifies that text modality contains much more semantic information. Besides, from a global perspective, as visualized in Fig.2, the dual attention and mutual-adjustment are crucial to ensure the effectiveness of intra-modal learning, also, the model's performance is greatly increased by both the intra-modal learning and the cross-modal interaction, which proves that our framework achieves a better joint cross-modal embedding space without losing the independence of uni-modality, except for the MELD dataset, the performance of speech-only intra-modal learning is not very significant.

3.3.3. Influence of the Hyper-Parameters

We attach three weights coefficients α, β, γ for our three classifiers with the constraint: $\alpha+\beta+\gamma=1$ with $0 \leq \alpha, \beta, \gamma \leq 1$ and we obtain $\alpha = 0.3$, $\beta = 0.3$, $\gamma = 0.4$ as the final weights. Furthermore, we conduct additional experiments of some representative hyperparameters to show how α, β, γ jointly affect the classification performance. We notice that the bi-modal classifier and text-only classifier contribute more to the final performance. Specifically, we evaluated the proposed framework with varied parameters $\alpha = [0.1, 0.2, \ldots, 0.5]$, $\beta = [0.1, 0.2, \ldots, 0.5]$, $\gamma = 1-\alpha-\beta$ while keeping other parameters as constant, and the results are shown in the right figure. The highest performance metrics (76% on WA and 77.84% on UA) are achieved at $\alpha = 0.3$, $\beta = 0.3$, $\gamma = 0.4$. Abnormal values of α and β (too large or too small) lead to degraded performance, indicating that the moderate weight of three classifiers is beneficial for obtaining the optimal late-fusion score.



(a) w/o DAB (b) w/o CMI (c) w/o MA (d) Final

Figure 2: *t-SNE visualization of feature distribution. (a)(b)(c) are the final MCFN without Dual Attention Block, Cross-Modal Interaction, Mutual-Adjustment, (d) is the final MCFN*

Table 3: *Ablation study results (%) on the proposed MCFN. "w/o DAB" means no temporal-channel wise dual-attention is exploited, "w/o CMI" means no cross-modal interaction is applied, "w/o MA" means the mutual adjustment is removed.*

	IEMOCAP		MELD	
Methods	WA(%)	UA(%)	Acc(%)	WF1(%)
Speech only	62.1	60.3	48.1	36.7
Text only	68.1	69.1	63.1	59.8
Bimodal	74.5	75.8	63.6	60.5
Ablation				
w/o DAB	75.1	75.8	64.2	61.4
w/o CMI	74.8	75.7	63.8	61.1
w/o Intra-modal^s	74.5	74.9	64.2	62.0
w/o Intra-modal^t	73.2	74.1	63.2	61.5
w/o MA	75.5	76.4	63.5	61.1
Proposed	**76.0**	**77.8**	**64.5**	**62.2**

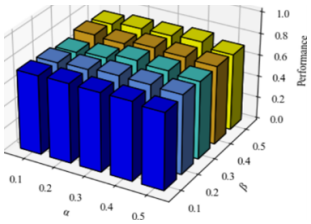


Figure 3: *Influence of three weights coefficients on IEMOCAP*

4. Conclusion

In this paper, we present a multimodal fusion framework for the emotion recognition. This model integrates the dual attention-based intra-modal learning with modality-collaborative learning, and achieves a hierarchical fusion so that both uni-modal and bi-modal can be interactively optimized without loss of the original modality-specific information. The experiments on the benchmark datasets show that our proposed method achieves competitive performance. Additionally, multimodal research with the visual information is left as our future work besides of the acoustic and semantic modalities.

5. References

[1] B. Li, D. Yaoitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *Proc. ICASSP. IEEE, 2019*, pp. 5876–5880.

[2] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 402–416, 2021.

[3] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. ICASSP. IEEE, 2021*, pp. 6269–6273.

[4] X. Zhang, Y. Zou, and W. Wang, "Ld-cnn: A lightweight dilated convolutional neural network for environmental sound classification," in *Proc. ICPR, 2018*, pp. 373–378.

[5] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP, 2014*, pp. 1532–1543.

[6] B. Kim, J. Seo, and M.-W. Koo, "Randomly wired network based on roberta and dialog history attention for response selection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2437–2442, 2021.

[7] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. ICASSP. IEEE, 2020*, pp. 6484–6488.

[8] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[9] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. ICASSP. IEEE, 2019*, pp. 2822–2826.

[10] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3592–3603, 2021.

[11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL, 2019*, p. 6558.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[13] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *Proc. INTERSPEECH, 2016*.

[14] 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.

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL, 2019*, pp. 527–536.

[16] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. ACL, 2018*, p. 2225.

[17] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *Proc. ICASSP. IEEE, 2022*, pp. 6897–6901.

[18] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proc. AAAI, 2019*, pp. 6818–6825.

[19] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. ACM MM, 2020*, pp. 2852–2861.

[20] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.