Context-aware Multimodal Fusion for Emotion Recognition

Conference Paper · September 2022

DOI: 10.21437/Interspeech.2022-10592

5 authors, including:

Jinchao Li
The Chinese University of Hong Kong
10 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Shuai Wang
Shanghai Jiao Tong University
66 PUBLICATIONS **1,388** CITATIONS
SEE PROFILE

Xunying Liu
The Chinese University of Hong Kong
278 PUBLICATIONS **6,207** CITATIONS
SEE PROFILE

Interspeech 2022
18-22 September 2022, Incheon, Korea

Context-aware Multimodal Fusion for Emotion Recognition

Jinchao Li[*1], *Shuai Wang*[2], *Yang Chao*[2], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong, China
[2]Lightspeed & Quantum Studios, Tencent, Shenzhen, China
[1]{jcli, xyliu, hmmeng}@se.cuhk.edu.hk, [2]{svsamwang, youngchao}@tencent.com

Abstract

Automatic emotion recognition (AER) is an inherently complex multimodal task that aims to automatically determine the emotional state of a given expression. Recent works have witnessed the benefits of upstream pretrained models in both audio and textual modalities for the AER task. However, efforts are still needed to effectively integrate features across multiple modalities, devoting due considerations to granularity mismatch and asynchrony in time steps. In this work, we first validate the effectiveness of the upstream models in a unimodal setup and empirically find that partial fine-tuning of the pretrained model in the feature space can significantly boost performance. Moreover, we take the context of the current sentence to model a more accurate emotional state. Based on the unimodal setups, we further propose several multimodal fusion methods to combine high-level features from the audio and text modalities. Experiments are carried out on the IEMOCAP dataset in a 4-category classification problem and compared with state-of-the-art methods in recent literature. Results show that the proposed models gave a superior performance of up to 84.45% and 80.36% weighted accuracy scores respectively in Session 5 and 5-fold cross-validation settings.

Index Terms: Emotion Recognition, multimodality, transfer learning, deep learning

1. Introduction

Automatic emotion recognition (AER) aims to determine the emotional state of a given expression automatically. It plays an essential role in various applications such as human-computer interactions and psychological assessments. [1–4]. AER is an inherently a complex multimodal task, since humans express and perceive emotions in different ways and across modalities, such as speech intonation [5, 6], linguistic content [7, 8], facial expression, etc. [9–11].

With recent advancements in Self-Supervised Learning (SSL), the emotion representations are shifting from handcrafted features, e.g., acoustic pitch and energy [12] or textual keywords and semantic information [13], to high-level embeddings extracted by pre-trained models, e.g. BERT and HuBERT [14, 15]. To further leverage these high-level features, various network architectures have been explored in the latest AER tasks [16].

Although these approaches have yielded respectable results, several key issues remain to be addressed. First, the emotion of an utterance is usually strongly related to the dialog context, yet most utterance-level feature modelling methods have not captured such information. Second, relying solely on either acoustic or linguistic information does not offer sufficient robustness in emotion recognition, which leads to increasing attention devoted to the use of multi-modal approaches. The heterogeneity across modalities calls for research in multimodal fusion strategies, including context-dependent fusion [17], contextual Long-Short Term Memory (LSTM) [18], co-attentional fusion [19], score fusion [20], dynamic convolution [21], self-supervised learning [22], time synchronous and asynchronous fusion [23], and etc. [24].

This work proposes a context-aware multimodal fusion framework for the AER task, consisting of a context-aware SSL-based feature extractor and Transformer-based audio-text fusion paradigms. We choose the pretrained WavLM [25] and BERT [26] models to encode the raw audio and text inputs into frame-level or token-level embeddings, respectively. Then, the extracted high-level features are calibrated and condensed with a Squeeze-and-Excitation (S&E) block [27], followed by a Fully-Connected (FC) Layer and Layer Normalization [28] which map the two modalities into the same space. After that, a context-aware Convolution block is adopted to enhance the target utterance with the context information for the text modality. We then proposed a novel Multimodal Transformer (MMT) fusion module to integrate the aforementioned features. The key component of the MMT is the directional pairwise cross-modal attention [29] based Transformer (CMT), which allows interactions between modalities with distinct time steps. After the fusion module, the integrated output is then fed into an emotion classifier.

We conduct the experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [30] dataset with a 4-category classification setup. The proposed models are evaluated with averaged weighted accuracy (WA) and unweighted accuracy (UA) in a leave-one-session-out (5-fold) cross-validation (CV) setting. Comparison is made with other state-of-the-art (SOTA) approaches published in recent literature. The rest of the paper is organized as follows. Section 2 introduces the proposed methods for uni- and multi-modal AER. Section 3 describes the experimental setups and results. Further analysis of the proposed methods and comparison with recent literature are given in Section 4. Finally, Section 5 concludes the paper and presents possible future research directions.

2. Proposed Approach

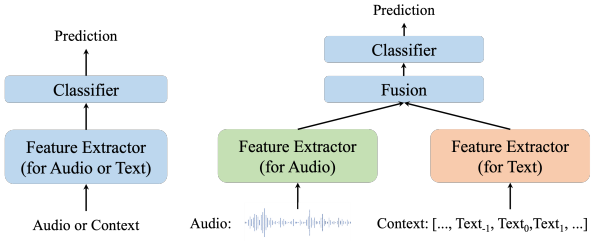
The proposed AER framework is illustrated in Figure 1. It is composed of three modules, a high-level feature extractor, a cross-modal fusion module (in the multimodal pipelines), and a classifier. In the following, we will introduce these three modules in detail.

2.1. High-level feature extractor

The recent success of large pre-trained models motivates this work to adopt novel, high-level features from self-supervised

*: Work done during the internship at Tencent Lightspeed & Quantum Studios

10.21437/Interspeech.2022-10592



(a) Unimodal (b) Multimodal

Figure 1: *Proposed AER frameworks: (a) unimodal and (b) multimodal.*

learning models [14, 15].

For the audio modality, we adopt WavLM [25], or more precisely, the WavLM-Large variant. It is trained on 94k hours of diverse data and contains a convolutional feature encoder and 24 stacked Transformer encoders. To avoid information leakage caused by only using the states from the last layer of the pre-trained model, we adopt the outputs of the feature encoder and all Transformer encoders as acoustic features. Since the WavLM model is trained using features of a 20ms stride, we can obtain a feature map with a shape of $50 \times 25 \times 1024$ (dimensions of time, layer and feature respectively) for each second of the audio input.

For the text modality, the popular BERT [26] is adopted, which contains a tokenizer and 12 Transformer encoders. Similar to the audio modality, we take the output from all layers, resulting in a feature map with a size of $1 \times 13 \times 768$ for each token of the input sentence.

As described above, the raw features extracted from the WavLM or BERT for the two modalities are large and redundant, to further condense the information which is helpful to the AER task, we propose two kinds of feature extractors as illustrated in Figure 2.

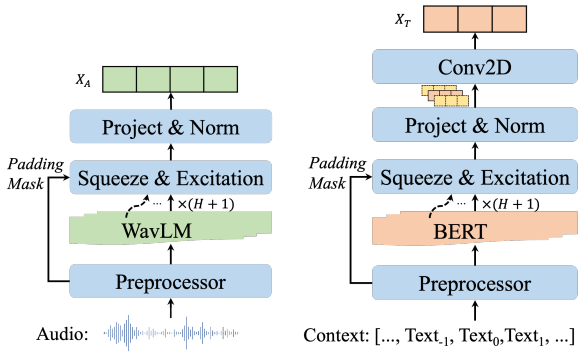
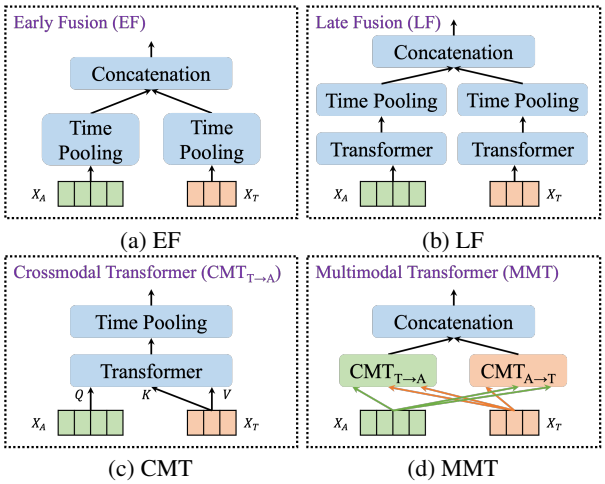


Figure 2: *Feature extractor (a) for audio and (b) for text.*

First, we adopt a trainable Squeeze-and-Excitation (S&E) block [27] to calibrate layer-wise feature responses adaptively, where the reduction ratio is simply the number of layers in the corresponding pretrained model here. Then, we project the calibrated features into the dimension of 16 to reduce feature redundancy while retaining the intra-class variability. The projector is an MLP with Layer-Normalization, which can also map audio and text modalities into the same dimensional space for fusion.

Moreover, considering that the emotion of each utterance is often related to its context in a spoken dialogue, we follow the method used in [23] to extract several embeddings from consecutive utterances in the textual modality. These embeddings are further aggregated into a context-aware textual embedding using a convolutional layer with the stride of 1 and kernel size of 3.

2.2. Cross-modal fusion



(a) EF (b) LF

(c) CMT (d) MMT

Figure 3: *Proposed fusion paradigms: (a) early fusion (EF), (b) late fusion (LF), (c) CMT, and (d) MMT.*

The projected and convoluted features are either fed into a classifier directly in the unimodal process, or fused first in the multimodal process. To combine the features from different modalities, we designed several fusion strategies, which are demonstrated in Figure 3.

Early Fusion (EF) denotes the simple concatenation of the temporal-pooled features from the audio and text feature extractors, while Late Fusion (LF) inserts an extra Transformer block to adjust the features further.

Then, we proposed the more elaborate cross-modal Transformer (CMT) module, which can be further used in the multimodal Transformer (MMT) fusion module. CMT has the same backbone as the vanilla Transformer [31], but the query and key-value pairs come from different modalities in the Attention block.

We consider two modalities, source s and target t, with two (potentially non-aligned) sequences in each of them denoted as $X_s \in \mathbb{R}^{T_s \times d_s}$ and $X_t \in \mathbb{R}^{T_t \times d_t}$, respectively. The notations $T_{(\cdot)}$ and $d_{(\cdot)}$ represent sequence length and feature dimension respectively. We define the queries as $Q_t = X_t W_{Q_t}$, keys as $K_s = X_s W_{K_s}$ and values as $V_s = X_s W_{V_s}$, where $W_{Q_t} \in \mathbb{R}^{d_t \times d_k}$, $W_{K_s} \in \mathbb{R}^{d_s \times d_k}$ and $W_{V_s} \in \mathbb{R}^{d_s \times d_v}$. Then the latent adaptation from source s to target t is presented as the cross-modal attention (CMA):

$$\begin{aligned} \mathrm{CMA}_{(s\to t)} &= \mathrm{softmax}(\frac{Q_t K_s^T}{\sqrt{d_k}}) V_s \\ &= \mathrm{softmax}(\frac{X_t W_{Q_t} W_{K_s}^T X_s^T}{\sqrt{d_k}}) X_s W_{V_s}. \end{aligned} \tag{1}$$

Based on the CMA block, the Cross-modal Transformer enables one modality to receive information from another. We fuse the audio and text modalities by viewing them as source modalities in CMT, respectively, namely CMT (audio $\to$ text) and CMT (text $\to$ audio). An average time pooling layer follows the CMT to compress variant time lengths into one. We

also propose a Multimodal Transformer (MMT) by concatenating the outputs of these two kinds of CMTs. The outputs of these fusion modules are finally fed into a fully-connected (FC) layer for emotion classification.

2.3. Classifier

For the multimodal systems, we adopted a simple MLP comprised of fully connected (FC) layers as the classifier for the fused features. For the uni-modal systems, the features obtained from the feature extractor are directly fed into the final emotion classifier. Besides the simple MLP classifier used in the multimodal systems, we experimented with a more complicated Transformer encoder (TE) architecture to validate whether the performance may be further improved to compare against the multimodal systems with transformer-based fusion modules.

2.4. Training & Fine-tuning

We adopt a cross-entropy loss function for the outputs from unimodal or multimodal pipelines. In the training process, we first freeze the pre-trained WavLM or BERT model and partially fine-tune them after training convergence. The fine-tuning process can be viewed as domain adaptation training. Instead of unfreezing the whole parameters in WavLM or BERT model, we only fine-tune the feature extractor part, i.e., CNN encoders in WavLM and input embeddings in BERT, to adapt the bottom-level feature space while preventing over-fitting.

3. Experiments

In this section, we introduce the experimental dataset, details of our experimental settings, and finally, results of the proposed models and comparisons with other state-of-the-art approaches.

3.1. Dataset

We utilize the IEMOCAP dataset to evaluate our models for the AER task. The dataset has approximately 12 hours of data and consists of scripted and improvised dialogues by 10 speakers in 5 sessions. To be consist with previous works, we use 4 emotional classes in this work, including "angry", "happy" (merged with "excited"), "sad", and "neutral". Thus, there are 5,531 utterances totally (1,103 "angry", 1,626 "happy", 1,084 "sad", and 1,708 "neutral").

We evaluate the models with a leave-one-session-out (5-fold) cross-validation (CV) setting, i.e., four sessions for training and validation while the remaining one for testing. Since the dataset is slightly imbalanced among emotion categories, the averaged weighted accuracy (WA) and unweighted accuracy (UA, balanced by class weights) are reported as metrics.

3.2. Experimental Setup

In the experiments, audios are clipped into 5 seconds and texts are clipped into 512 tokens by trimming or padding. To be consistent with previous work in [23], we trim the voiced 5 seconds of audio and 512 words at the beginning for long utterances, where the emotions are supposed to be expressed intensely. Consider the context effect reported in [23], the number of textual contexts we select is 9, i.e., 4 utterances before and 4 after the target utterance are selected in sequence.

The systems are trained by using AdamW [32] optimizer with a learning rate of 10^{-3} during initial training and 10^{-5} during fine-tuning, along with a weight decay of 10^{-3}. A dropout layer (0.5) between every two modules is used for regularization throughout our training. The batch size for training is 32. Moreover, we randomly select one speaker as a development set and monitor its loss for fine-tuning and early stopping. If the loss has not decreased for 5 epochs, the fine-tuning process starts; if the loss has not decreased for 10 epochs, the training stops. The last model is selected and then evaluated.

3.3. Results

We present the results with both unimodal and multimodal pipelines in Table 1.

Sys.	Model	Modality	WA(%)	UA(%)
1	WavLM+FC❄	A	66.15	65.35
2	WavLM+FC☼	A	66.70	66.71
3	WavLM+TE❄	A	63.51	64.67
4	WavLM+TE☼	A	**67.99**	**68.24**
5	BERT+FC❄	T	74.33	76.57
6	BERT+FC☼	T	**77.67**	**78.76**
7	BERT+TE❄	T	74.55	76.41
8	BERT+TE☼	T	76.09	76.57
9	$\text{CMT}_{(A\to T)}$ ❄	A + T	67.76	69.37
10	$\text{CMT}_{(A\to T)}$ ☼	A + T	68.07	67.93
11	$\text{CMT}_{(T\to A)}$ ❄	A + T	74.98	76.85
12	$\text{CMT}_{(T\to A)}$ ☼	A + T	**77.75**	**78.60**
13	EF (1 $\oplus$ 5) ❄	A + T	77.61	77.85
14	EF (2 $\oplus$ 6) ☼	A + T	79.80	80.65
15	LF (3 $\oplus$ 7) ❄	A + T	79.91	**81.88**
16	LF (4 $\oplus$ 8) ☼	A + T	79.91	81.15
17	MMT (9 $\oplus$ 11) ❄	A + T	78.58	80.72
18	MMT (10 $\oplus$ 12) ☼	A + T	**80.36**	**81.70**

Table 1: *5-fold CV mean results of unimodal and multimodal systems on IEMOCAP dataset. "A" and "T" refer to audio and text modalities respectively. "$a \oplus b$" means the system can be viewed as a joint of system a and system b.*
❄: frozen, ☼: finetuned.

3.3.1. Unimodal Results

As shown in Table 1, the textual modality (Sys. 5-8) outperforms the audio modality (Sys. 1-4) by a large margin on the IEMOCAP dataset. This fact aligns with the observations in prior work [33]. We can also find that the fine-tuned features generally outperform frozen features.

3.3.2. Multimodal Results

Table 1 also shows the effectiveness of the proposed fusion models. First, we consider the performance of Crossmodal Transformers (Sys. 9-12). It shows that $\text{CMT}_{(T\to A)}$ outperforms $\text{CMT}_{(A\to T)}$, which also implies the superiority of the textual modality, as seen in the unimodal results. Then, we compare the Cross-modal Transformers with unimodal systems (Sys. 1-8) and find that $\text{CMT}_{(T\to A)}$ slightly improves over the textual performance (Sys. 11-12 v.s. Sys. 5-8). Finally, we consider the performance of the fusion of unimodal MLPs (Sys. 13-14), vanilla Transformers (Sys. 15-16), and Cross-modal Transformers (Sys. 17-18). We find that the fusion of CMT models is best among other systems, implying that the Cross-modal Transformer-based interaction benefits the AER task.

4. Discussion

4.1. Comparison with previous literature

In Table. 2, we compare the proposed models with existing state-of-the-art systems published on the IEMOCAP dataset using audio and textual modalities. The state-of-the-art fusion methods listed here including score fusion [20], Co-Attentional Fusion [19], and time-synchronous and asynchronous concatenation [23].

Modality	Test setting	WA(%)	UA(%)
A + T❄ [23]	Session 5	83.08	83.22
A + T❄ (ours)	Session 5	78.32	80.56
A + T☼ (ours)	Session 5	**84.45**	**84.39**
A + T☼ [19]	5-fold CV	-	75.46
A + T☼ [20]	5-fold CV	73.5	73.0
A + T❄ [23]	5-fold CV	77.57	78.41
A + T❄ (ours)	5-fold CV	79.91	**81.88**
A + T☼ (ours)	5-fold CV	**80.36**	**81.70**

Table 2: *Comparative results of the proposed systems with baseline state-of-the-art approaches on IEMOCAP dataset using audio ("A") and text ("T") modalities.*
❄: frozen, ☼: finetuned.

We also report results on a single Session 5 subset comparable to previous literature. Here, we do not include other works that do not focus on 4-category classification, or are simply tested with a random CV setting without considering the speaker overlap or data leakage.

As shown in Table 2, the proposed models achieve best performance, reaching 84.45% WA (84.39% UA) on Session 5, and 80.36% WA (81.70% UA) on 5-fold CV for the IEMOCAP dataset.

4.2. Analysis

We conducted several studies on Session 5 of the IEMOCAP dataset to analyze the performance of the proposed systems.

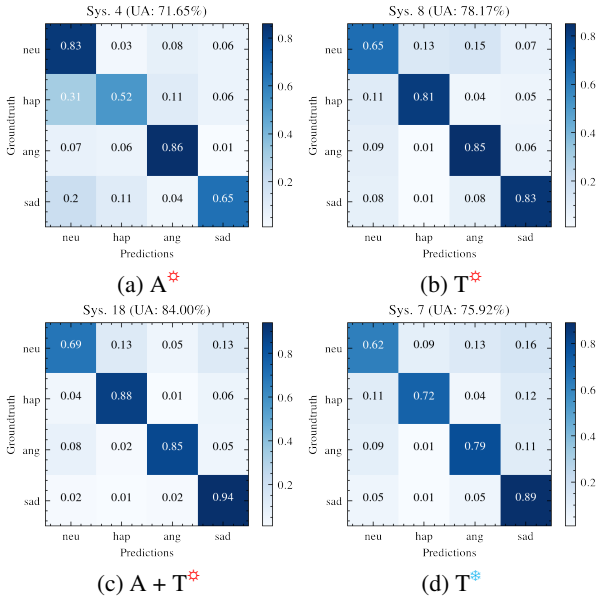
4.2.1. Frozen v.s. Fine-tuned

From both unimodal and multimodal results in Table 1, we can find that the fine-tuned systems generally outperform the frozen systems. The confusion matrices in Figure 4 also shows the improvement from the frozen (Figure 4d) to fine-tuned (Figure 4b) state of the same model.

4.2.2. Unimodal v.s. Multimodal

From the results in Table 1, we can also find that the multimodal systems generally outperform the unimodal systems. The confusion matrices of uni- and multimodal systems are illustrated in Figure 4, where Sys. 4, Sys. 8, and Sys. 18 represent audio-only, text-only and multimodal systems respectively. Several interesting observations were found in Figure 4.

First, the audio model (Figure 4a) performs well on "neutral" and "angry" emotions, but poorly on "happy" and "sad" emotions, most of which are falsely predicted as "neutral". An intuitive reason is that speech intonations of "neutral" are more distinguishable with "angry", but may confuse with "happy" and "sad" emotions. Second, the textual model (Figure 4b) performs well on "happy", "angry" and "sad", but poorly on "neutral". This is similar to the multimodal system (Figure 4c). A



(a) A☼ (b) T☼

(c) A + T☼ (d) T❄

Figure 4: *The normalized confusion matrices of unimodal and multimodal systems evaluated on Session 5 of IEMOCAP. "A" and "T" refer to audio and text modalities respectively. "neu", "hap" and "ang" refer to "neutral", "happy" and "angry" respectively.*
❄: frozen, ☼: finetuned.

possible reason is that there are some emotional keywords in a "neutral" sentence, e.g., "Why does that **bother** you?". The emotional words appearing in a "neutral" sentence may bring confusion to the model for the textual modality. Finally, we can observe that the multimodal system (Figure 4c) achieves the best performance and especially, surpass the accuracy of every emotion class compared with the text-only model (Fig. 4b), which reflects the synergetic attributes among the audio and text modalities.

5. Conclusion

In this paper, we propose a context-aware feature extractor and several multimodal fusion methods for the AER task. The SSL embeddings from context utterances with further bottom-level fine-tuning show the power of emotional representations. Experiments are carried out on IEMOCAP. The proposed multimodal Transformer-based method achieves 80.36% WA and 81.70% UA under the 5-fold CV setting, 84.35% WA and 84.39% UA on the Session 5, which shows the attainment of state-of-the-art performance on this dataset. In future work, we will investigate the robustness and generalization of the AER task in the cross-lingual or few-shot speaker adaptation settings.

6. Acknowledgements

The authors would like to thank Dr. Chao Zhang and Wen Wu for their helpful discussion on contextual modeling and dataset splitting. This project is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N).

7. References

[1] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 3687–3691.

[2] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[4] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[5] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[6] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[7] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," *arXiv preprint arXiv:1906.01267*, 2019.

[8] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. ICMI*, 2004, pp. 205–211.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[11] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI*, vol. 34, no. 02, 2020, pp. 1359–1367.

[12] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[13] Z.-J. Chuang and C.-H. Wu, "Emotion recognition from textual input using an emotional semantic network," in *Proc. ICSLP*, 2002.

[14] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in *Proc. ICCWAMTIP*. IEEE, 2020, pp. 117–121.

[15] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[16] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[17] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[18] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-based dense lstm for speech emotion recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 102, no. 7, pp. 1426–1429, 2019.

[19] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," in *Proc. INTERSPEECH*. IEEE, 2020.

[20] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," *arXiv preprint arXiv:2111.10202*, 2021.

[21] H. Wen, S. You, and Y. Fu, "Cross-modal dynamic convolution for multi-modal emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103178, 2021.

[22] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. SLT*. IEEE, 2021.

[23] 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.

[24] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] 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*, vol. 2019. NIH Public Access, 2019, p. 6558.

[30] 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.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[33] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 6892–6899.