INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Leveraging Label Information for Multimodal Emotion Recognition

*Peiying Wang, Sunlu Zeng, Junqing Chen, Lu Fan, Meng Chen*, Youzheng Wu, Xiaodong He*

JD AI Research, Beijing, China

{wangpeiying3,zengsunlu1,chenjunqing3,fanlu,chenmeng20,wuyouzheng1,xiaodong.he}@jd.com

## Abstract

Multimodal emotion recognition (MER) aims to detect the emotional status of a given expression by combining the speech and text information. Intuitively, label information should be capable of helping the model locate the salient tokens/frames relevant to the specific emotion, which finally facilitates the MER task. Inspired by this, we propose a novel approach for MER by leveraging label information. Specifically, we first obtain the representative label embeddings for both text and speech modalities, then learn the label-enhanced text/speech representations for each utterance via *label-token* and *label-frame* interactions. Finally, we devise a novel label-guided attentive fusion module to fuse the label-aware text and speech representations for emotion classification. Extensive experiments were conducted on the public IEMOCAP dataset, and experimental results demonstrate that our proposed approach outperforms existing baselines and achieves new state-of-the-art performance.
**Index Terms**: Multimodal emotion recognition, label embedding, cross-attention

## 1. Introduction

Generally, the emotion of spoken language is beyond the linguistic content of the utterance itself, and it is also related to the speaker's voice characteristics. To completely understand the emotion of the speaker, the text- and speech-based multimodal emotion recognition (MER) task was proposed to identify the emotion within an utterance [1].

Recently, MER has attracted more and more attention. In the early phase, most works explored rule-based and neural network-based methods [2, 3]. With the rapid development of self-supervised learning and pre-training, researchers attempt to tackle this task based on pre-trained models, e.g. BERT [4] and wav2vec2.0 [5]. For instance, Li et al. [6] proposed a context-aware multimodal fusion framework for the MER task, which applied BERT and WavLM as encoders. Chen et al. [7] proposed a key-sparse Transformer based on the RoBERTa and Wav2vec, which focuses more on emotion-related information. Despite their success, most of them only take labels as supervised signals while neglecting their inherent semantic information. Intuitively, the label information should be capable of helping the model to better understand the utterance. As shown in Figure 1, for the text input, the token "mad" is similar to the label *angry* in semantics. As to the speech input, some frame segments also have in common with the tonal label. Based on above observation, we argue that the model may be able to locate the task-oriented salient tokens/frames accurately under the guidance of the label information. Then, the model can pay more attention to the key information and effectively ignore the interference of redundant information. Therefore, as a kind of prior knowledge, leveraging label information is essential for the MER task.

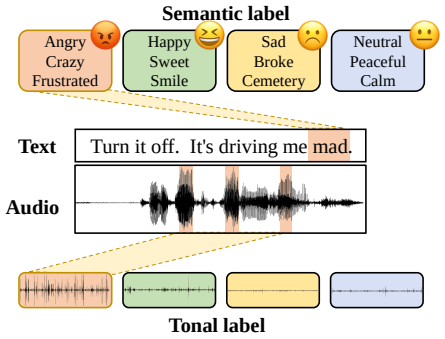


Figure 1: *Visualization of labels. The **semantic label** presents the emotion relevant words for each class, and the **tonal label** displays the waveforms generated by concatenating the key-frames under each class.*

Label embedding is to learn the embeddings of the labels in classification tasks and has been proven to be effective in computer vision and natural language processing [8–11], which enjoys a built-in ability to leverage alternative sources of information related to labels, such as class hierarchies or textual descriptions. However, there are rare speech-related work devoted to this technology. Take MER for example, there exist at least two obstacles that need to resolve. Firstly, labels are usually in the form of text. Due to the inherent disparities between the speech and the text, they cannot be directly exploited in speech-related tasks. How to obtain representative label embeddings for the speech modality becomes a big challenge. Secondly, when introducing the label information, it will increase the difficulty of multimodal fusion. How to project the text/speech representations into the same label embedding space and fuse the multimodal features seamlessly is also a critical issue.

In this work, we propose a novel framework for MER to tackle the above challenges. We first summarize representative tokens/frames from the training set for each class as their descriptions. For text, we extract salient words for each label based on the frequency of tokens. As for speech, we utilize wav2vec2.0 to discretize the whole dataset, and then extract salient frames from them. After obtaining that, we further devise a novel label-enhanced multimodal emotion recognition model (LE-MER). Formally, given an utterance and the extracted label information, we first adopt BERT and wav2vec2.0 to learn representations for the text and speech input. To locate the salient tokens/frames in the utterance, we conduct label-text/speech interactions by introducing a *label-token* at-

*Corresponding author.


10.21437/Interspeech.2023-1732

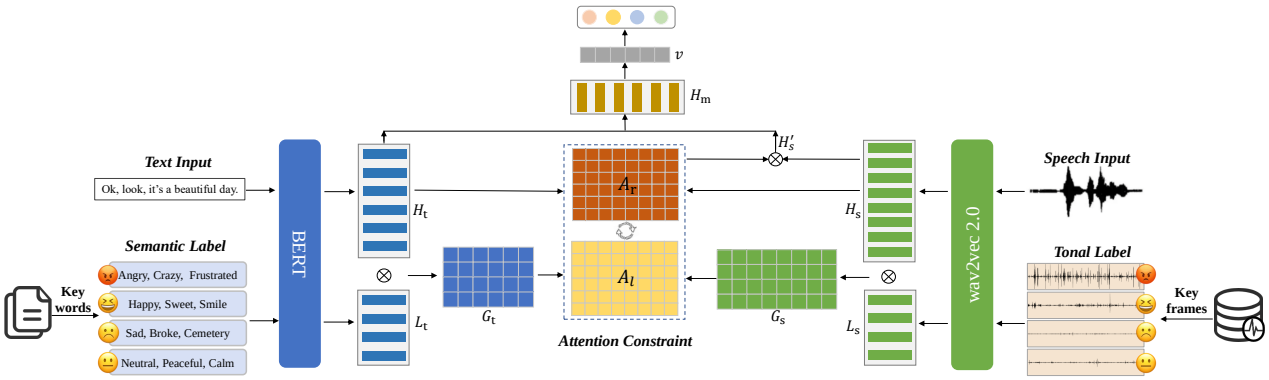


Figure 2: *The architecture of our proposed model LE-MER.*

tention mechanism for the text and a *label-frame* one for the speech, which encourages the model to pay more attention to the emotion-related tokens/frames. Based on the above two cross-attention maps, we further introduce a label-guided attention mechanism to fuse the text and the speech. Since the inputs of this mechanism involve emotion-related information, it is capable of aligning the text and the speech from the emotional perspective.

Our main contributions are summarized as follows: 1) To the best of our knowledge, this is the first work exploring a label embedding enhanced model for speech emotion recognition. 2) We propose a novel label-guided cross-attention mechanism to fuse different modalities, which is capable of learning the alignment between speech and text from the perspective of emotional space. 3) We show the effectiveness of our method on the IEMOCAP dataset with significant improvements compared with the baseline methods.

## 2. Proposed Approach

The overview architecture of our proposed LE-MER is illustrated in Figure 2. LE-MER consists of three modules: a semantic label enhanced text encoder, a tonal label enhanced speech encoder, and a multi-modal fusion module with label-guided cross attention.

### 2.1. Semantic-label enhanced text encoder

Inspired by the success of Pre-trained Language Model (PLM) [9, 12, 13] on numerous NLP tasks, we apply the BERT [4] as the text encoder without loss of generality. For each utterance $u_t$, we feed it into the BERT and get the sequence representation $\mathbf{H}_t \in \mathbb{R}^{l_t \times d_t}$, where $l_t$ is the length of utterance and $d_t$ is the dimension.

In order to get the emotion-aware text representation, we fuse the semantic label information into the text encoder. Firstly, we extract the keywords on the text corpus under one class in the training set to get representative textual label descriptions. Specifically, we adopt the commonly used TF-IDF [14] algorithm to extract the Top-K words, then we feed these label descriptions into BERT, and obtain the semantic label embedding $\mathbf{L}_t \in \mathbb{R}^{c \times d_t}$ by averaging the token embeddings of all label descriptions, where $c$ denotes the number of classes.

After embedding both the words and the labels into a joint space, we can obtain the *label-token* attention matrix $\mathbf{G}_t \in \mathbb{R}^{l_t \times c}$ by computing the cosine similarity between the text representation and label embedding as follows:

$$\mathbf{G}_t = \frac{\mathbf{H}_t \cdot \mathbf{L}_t^T}{\|\mathbf{H}_t\|_2 \|\mathbf{L}_t\|_2} \tag{1}$$

Then we introduced a new objective based on the *label-token* interaction to encourage the emotion relevant words to be weighted higher than the irrelevant ones. Specifically, we conduct mean-pooling on the attention matrix $\mathbf{G}_t$ along the axis of sequence length, which is used as the discriminator for each class to judge the emotional relevance. Finally, we obtain the predicted logits $\mathbf{p}_g^t$ and the loss $\mathcal{L}_g^t$:

$$\mathbf{p}_g^t = \text{Softmax}(\text{Meanpooling}(\mathbf{G}_t)) \tag{2}$$

$$\mathcal{L}_g^t = \text{CE}(\mathbf{y}, \mathbf{p}_g^t) \tag{3}$$

where $\text{CE}(\cdot, \cdot)$ refers to the cross entropy loss.

### 2.2. Tonal-label enhanced speech encoder

The recent success of large pre-trained models [15–22] motivates us to adopt novel, high-level features from self-supervised learning models. For the audio modality, we use wav2vec2.0 [5] as our speech encoder. For each waveform of utterance, we obtain a sequence of contextualized representations from the output of wav2vec2.0 $\mathbf{H}_s \in \mathbb{R}^{l_s \times d_s}$, where $l_s$ is the number of time frames, and $d_s$ is the feature dimension.

Similarly, we leverage the label information to obtain the emotion-aware speech representation. In order to represent the tonal label, we adopt a unified method as in text label embedding. That is, the key audio frames which contain representative tonal information will be extracted to generate the speech label embeddings for each class. To this end, we need to obtain the discrete representations of speech first. The quantizer module of the pre-trained wav2vec2.0 discretizes the output of the CNN feature encoder into a finite set [5], enabling the application of key-frames extraction on the speech data. In the same way, the TF-IDF is adopted to select the Top-K emotion-relevant frames under the same class. The embeddings of these emotion-relevant frames extracted from (codebook of) wav2vec2.0 will be averaged to produce the final tonal label embeddings $\mathbf{L}_s \in \mathbb{R}^{d_s \times c}$, where $d_s$ is the feature dimension identical to the dimension of $\mathbf{H}_s$.

Through the above way, we have embedded both the speech and the tonal label into a shared latent space, and then the *label-frame* interaction matrix can be computed in the following way:

$$\mathbf{G}_s = \frac{\mathbf{H}_s \cdot \mathbf{L}_s^T}{\|\mathbf{H}_s\|_2 \|\mathbf{L}_s\|_2} \tag{4}$$

where $\mathbf{G}_s \in \mathbb{R}^{l_s \times c}$, and each element indicates the similarity between the frames and the emotion category. In order to figure out the emotion related frames with the guidance of tonal label, we introduced another objective based on *label-frame* interaction. Specifically, the mean-pooling is conducted on the

interaction matrix $\mathbf{G}_s$ along the frame axis to aggregate utterance level emotional correlation score.

$$\mathbf{p}_g^s = \text{Softmax}(\text{Meanpooling}(\mathbf{G}_s)) \tag{5}$$

The *label-frame* interaction based objective can be written as:

$$\mathcal{L}_g^s = \text{CE}(\mathbf{y}, \mathbf{p}_g^s) \tag{6}$$

which directly encourages the speech encoder to pay more attention to the emotional frames.

### 2.3. Multi-modal fusion with label-guided cross attention

Multimodal fusion technology for speech emotion recognition has been widely studied in recent years, including cross-attention fusion [23], co-attention fusion [24], score fusion [25], time synchronous and asynchronous fusion [26], multimodal transformer [6], and etc. However, all these fusion mechanisms just devoted to aggregate word and speech embeddings, while ignored the the rich prior information contained in the emotion labels. We argue that the emotion labels can serve as a guidance to integrate the two modalities more efficiently. To this end, we elaborate a novel *label-guided cross-attention* to fuse multimodal emotion-related information.

The cross-attention mechanism have been proposed to capture the fine-grained interactions between the hidden representations of tokens and frames [23, 27]:

$$\mathbf{A}_r = \text{Softmax}(\mathbf{H}_t \mathbf{H}_s^T \mathbf{W}) \tag{7}$$

where $\mathbf{A}_r \in \mathbb{R}^{l_t \times l_s}$, and $\mathbf{W} \in \mathbb{R}^{d_s \times d_t}$ . Next, we can obtain the aligned hidden audio representation $\mathbf{H}_s'$ by weighting $\mathbf{H}_s$ with cross-attention $\mathbf{A}_r$, and the multimodal features $\mathbf{H}_m$ can be obtained by concatenating the text representation and the aligned speech representation:

$$\mathbf{H}_s' = \mathbf{A}_r \mathbf{H}_s^T, \mathbf{H}_m = [\mathbf{H}_t, \mathbf{H}_s'] \tag{8}$$

Considering that the text and speech have been projected into the target emotional space in sections 2.1 and 2.2, we directly multiply the *label-token* interaction $\mathbf{G}_t$ and the *label-frame* interaction $\mathbf{G}_s$ to obtain the label-guided cross-attention matrix:

$$\mathbf{A}_l = \mathbf{G}_t \cdot \mathbf{G}_s^T \tag{9}$$

where $\mathbf{A}_l \in \mathbb{R}^{l_t \times l_s}$, and each element indicates the similarity between text tokens and speech frames from the perspective of emotional correlation. Compared with $\mathbf{A}_l$, $\mathbf{A}_r$ solely represents the inherent semantic relations between the text and the speech, instead of emotion specific. To bridge the gap and integrate the emotion-aware relations into $\mathbf{A}_r$, we propose an *Attention Constraint Module* which adopts $\mathbf{A}_l$ to guide $\mathbf{A}_r$, and we implement it with the mean squared error as follows:

$$\mathcal{L}_c = ||\mathbf{A}_l - \mathbf{A}_r||^2 \tag{10}$$

Finally, we aggregate the emotion-aware multimodal features $\mathbf{H}_m$ into a fixed-length vector $\mathbf{v}$ via max-pooling operation, and then feed $\mathbf{v}$ into a linear projection to obtain the prediction result and optimize it with the cross-entropy loss:

$$\mathcal{L}_m = \text{CE}(\mathbf{y}, \text{Softmax}(\text{Linear}(\mathbf{v}))) \tag{11}$$

The overall loss function of the LE-MER is summarized as follows:

$$\mathcal{L} = \mu_1 \mathcal{L}_m + \mu_2 \mathcal{L}_c + \mu_3 \mathcal{L}_g^t + \mu_4 \mathcal{L}_g^s \tag{12}$$

where $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ are hyperparameters.

# 3. Experiments

In this section, we present the dataset, the results compared with other state-of-the-art approaches and the related analysis.

### 3.1. Dataset

We evaluate our proposed model on the commonly used Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [1], which contains approximately 12 hours of audiovisual data. Among them, the text transcriptions, along with the corresponding audio, consist of five dyadic sessions where actors perform improvisations or scripted scenarios. To be consist with previous works [6, 7, 26, 28–30], we conduct experiments on 5531 utterances from four categories: *angry*, *happy* (merged with *excited*), *sad*, and *neutral*. We evaluate the model by a leave-one-session-out (5-fold) cross-validation (CV) strategy and adopt the average weighted accuracy (WA) and unweighted accuracy (UA) as evaluation metrics.

### 3.2. Experimental Setup

**Data preprocessing.** For speech modality, the 80-dimensional Log Mel-spectrograms [31] of each speech waveform are extracted by a 25ms window size with a 10ms step size and then normalized to the standardized normal distribution in utterance level. SpecAugment [32] is also applied to the extracted acoustic feature to improve the generalization ability of the model. For text modality, we use historical utterances to enhance performance, as they can provide contextual information as well as some additional clues to the current utterance [33]. Specifically, no more than ten historical utterances are spliced for each utterance and the maximum token length is limited to 150.

**Settings.** The pre-trained wav2vec2.0-conformer-BASE [34] and BERT-BASE [4] model are employed as our speech and text encoder, respectively. Following [18, 35–37], wav2vec2.0 is pre-trained on 960h LibriSpeech dataset. In addition, 2nd stage pretraining is applied to the pre-trained wav2vec2.0-conformer on the training set. We adopt Adam as our optimizer with a warm-up of 8000 steps and set the learning rate to $10^{-6}$ for the wav2vec2.0 and $5 \times 10^{-6}$ for BERT model, while the batch size for training is 16. As for hyperparameters in Eq (12), we set $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ to 1, 0.5, 0.2, and 0.2 empirically[1].

**Baselines.** We compare our proposed LE-MER with several baselines: [6, 7, 28] adopted cross-attention mechanism to fuse multimodal information, where a modified key-sparse attention is proposed in [7]. Hou et al. [29] proposed a self-guided modality calibration network to achieve alignment between audio and text modalities. Wu et al. [26] proposed a two-branch neural network to capture correlations between multimodalitly from both word level and utterance level. Santoso et al. [30] proposed to use the combination of a self-attention mechanism and a word-level confidence measure (CM) to mitigate the errors in MER produced by ASR system.

### 3.3. Main Results

**Unimodal Results.** As we can observe from Table 1, the performance of the text encoder improves significantly after integrating historical utterances (A2), proving that historical utterances can provide additional cues to support the current utterances. In addition, we also investigate the effects of different initialization for text label embeddings (A3-A5). We can observe that

[1]The source code will be publicly available at: `https://github.com/Digimonseeker/LE-MER`.

Table 1: *Comparison of our unimodal results on IEMOCAP dataset where "LE" denotes label embedding.*

| Sys. | Model | WA(%) | UA(%) |
|---|---|---|---|
| A1 | BERT | 67.34 | 67.66 |
| A2 | + historical utterances | 77.46 | 78.38 |
| A3 | + historical utterances + LE (random init) | 77.51 | 78.52 |
| A4 | + historical utterances + LE (label words init) | 78.03 | 78.88 |
| A5 | + historical utterances + LE (TF-IDF init) | **78.11** | **78.92** |
| B1 | wav2vec2.0 | 73.92 | 74.48 |
| B2 | + 2nd stage | 75.73 | 76.44 |
| B3 | + 2nd stage + LE (random init) | 76.20 | 76.80 |
| B4 | + 2nd stage + LE (BERT embedding init) | 76.48 | 77.14 |
| B5 | + 2nd stage + LE (codebook init) | **76.74** | **77.74** |

A4 achieves superior performance than A3, while the best results are achieved by A5, which substantiates that our keyword initialization scheme yields a more effective and representative label embeddings than the others. For speech modality, 2nd stage pretraining before finetuning can improve WA and UA by at least absolute 1.8 percent (B2 vs. B1). Moreover, speech label embeddings with all three types of initialization (B3-B5) bring varying contributions compared to B2. Some potential prior information from BERT embedding (B4) boosts the performance compared with B3. By getting rid of the shackles of the modality gap and benefiting from the discretized tonal label generated from pre-trained quantizer of wav2vec2.0, B5 makes a further improvement (B5 vs. B4) and achieves the best result.

**Multimodal Results.** In Table 2, we compare our multimodal results on the IEMOCAP dataset with the existing state-of-the-art methods, which share the same setting of data preprocessing with us, for a fair comparison. It shows that our proposed approach achieves state-of-the-art results compared to the others in terms of both WA and UA. Here we also present the result obtained by the score fusion scheme, which simply sums the logits produced by two unimodal models for predictions. This straightforward scheme can achieve favorable result that outperforms all the baselines, indicating that our label embeddings can facilitate unimodal encoders to locate the salient tokens/frames relevant to the specific emotion, thus yielding better result. Compared with methods based on attention mechanism, such as [6, 7], our model achieves superior performance, which proves that our proposed label-guided attentive fusion module can serve as a bridge to leverage cues from multimodality to integrate emotional information more effectively.

Table 2: *Comparison of our multimodal results with previous works on IEMOCAP dataset.*

| Model | WA(%) | UA(%) |
|---|---|---|
| Chen et al. [7] | 74.30 | 75.30 |
| Chen et al. [28] | 74.92 | 76.64 |
| Hou et al. [29] | 75.60 | 77.60 |
| Wu et al. [26] | 77.57 | 78.41 |
| Santoso et al. [30] | 78.40 | 78.60 |
| Li et al. [6] | 80.36 | 81.70 |
| Our Score Fusion | 81.32 | 82.18 |
| Ours | **82.40** | **83.11** |

### 3.4. Discussion

**Hyper-parameter Tuning of $K$.** To explore the optimal number of frames for the tonal label, we conduct a grid search to obtain its value. As shown in Figure 3, when $K$ is larger than 100, both WA and UA decrease to some extent, implying that label embeddings with larger $K$ contain some redundant information that is irrelevant with the corresponding type of emotion. Vice versa, label embeddings with smaller $K$ lack enough emotion-related information, causing performance degradation. Therefore, we set $K$ to 100 for the tonal label. As for the semantic label, we explore the optimal $K$ with the same method, and the best $K$ is set to 9. For sake of repetition, we omit the process here.

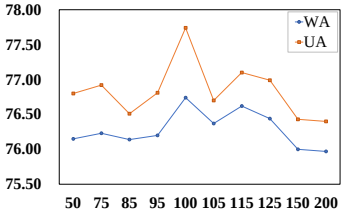


Figure 3: *Effect of K for speech label embeddings initialization.*

**Ablation study of Attention Constraint.** In this section, we further explore how to utilize the label-guided attention matrix to improve the model performance in terms of modality fusion, and we present the corresponding results in Table 3. A subtle decrease can be observed from C1 to C2 in terms of UA, revealing the superiority of the attention constraint scheme against simple summation of label-guided attention $\mathbf{A}_l$ and vanilla attention $\mathbf{A}_r$. We can attribute this performance gap to adopting attention constraint, as it provides a more delicate way to craft multimodal features $\mathbf{H}_m$ to be label-aware under the supervision of $\mathbf{A}_l$. Furthermore, we perform multimodal fusion with only $\mathbf{A}_l$ (C3) or $\mathbf{A}_r$ (C4). Further performance degradation can be observed by comparing either C3 or C4 with C2, validating the necessity of the interaction between $\mathbf{A}_l$ and $\mathbf{A}_r$.

Table 3: *Results of comparison between different fusion methods utilizing label-guided attention $\mathbf{A}_l$ and vanilla attention $\mathbf{A}_r$*

| Sys. | Model | WA(%) | UA(%) |
|---|---|---|---|
| C1 | Attention Constraint | **82.40** | **83.11** |
| C2 | $\mathbf{A}_r + \mathbf{A}_l$ | 82.39 | 82.75 |
| C3 | only $\mathbf{A}_l$ | 81.29 | 81.37 |
| C4 | only $\mathbf{A}_r$ | 81.08 | 81.68 |

**Attention Visualization.** To demonstrate the effectiveness of our unimodal label embeddings, we perform the visualizations of both $\mathbf{G}_s$ and $\mathbf{G}_t$ on one utterance in IEMOCAP and present them in Figure 4. In this example, waveform is aligned with the tokenized text, and both $\mathbf{G}_s$ and $\mathbf{G}_t$ have been averaged over the class dimension to generate the vectors $\widetilde{\mathbf{G}}_s$ and $\widetilde{\mathbf{G}}_t$. Results show that the $\widetilde{\mathbf{G}}_s$ assigns larger attention weights to emotion-related speech segments, while $\widetilde{\mathbf{G}}_t$ has higher weights on some emotional words, such as "cool". This reveals that our unimodal label embeddings can effectively guide the encoders to focus on emotion relevant information from the input.

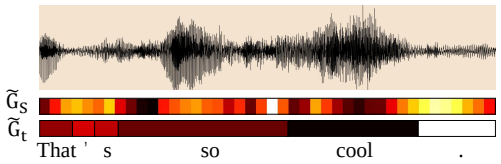


Figure 4: *Visualization of $\widetilde{\mathbf{G}}_s$ and $\widetilde{\mathbf{G}}_t$.*

## 4. Conclusions

In this paper, we presented LE-MER, a novel multimodal fusion framework for speech emotion recognition, which takes advantage of the both the textual and speech label information to extract the emotional token and frames, respectively. By mapping the speech and text representations to a common emotional space, we can learn the alignment between the text words and speech frames and fuse the emotional information more efficiently. Experimental results on the public IEMOCAP dataset demonstrated the superior performance of LE-MER and the importance of each component. In the future, we will explore how to extend this method to other speech tasks.

# 5. References


[1] 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, pp. 335–359, 2008.

[2] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *SLT 2018, Athens, Greece, December 18-21, 2018*. IEEE, 2018, pp. 112–118.

[3] R. Peri, S. Parthasarathy, C. Bradshaw, and S. Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," in *ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 6344–6348.

[4] 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.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] J. Li, S. Wang, Y. Chao, X. Liu, and H. Meng, "Context-aware multimodal fusion for emotion recognition," *Proc. Interspeech 2022*, pp. 2013–2017, 2022.

[7] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[8] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," *arXiv preprint arXiv:1805.04174*, 2018.

[9] Y. Xiong, Y. Feng, H. Wu, H. Kamigaito, and M. Okumura, "Fusing label embedding into bert: An efficient improvement for text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1743–1750.

[10] Z. Zhang, Y. Zhao, M. Chen, and X. He, "Label anchored contrastive learning for language understanding," *arXiv preprint arXiv:2205.10227*, 2022.

[11] Y. Le, Y. Zhao, M. Chen, Z. Quan, X. He, and K. Li, "Legal charge prediction via bilinear attention network," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1024–1033.

[12] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[13] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," *arXiv preprint arXiv:1912.02610*, 2019.

[14] S. Beliga, "Keyword extraction: a review of methods and approaches," *University of Rijeka, Department of Informatics, Rijeka*, vol. 1, no. 9, 2014.

[15] 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," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[17] L. Fu, S. Li, Q. Li, L. Deng, F. Li, L. Fan, M. Chen, and X. He, "Ufo2: A unified pre-training framework for online and offline speech recognition," in *ICASSP 2022-2023*, 2023.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[19] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[20] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[21] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*.

[22] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

[23] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[24] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[25] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," in *(ASRU) 2021*. IEEE, 2021, pp. 350–357.

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

[27] Y. Zhu, Z. Wang, H. Liu, P. Wang, M. Feng, M. Chen, and X. He, "Cross-modal transfer learning via multi-grained alignment for end-to-end spoken language understanding," *Proc. Interspeech 2022*, pp. 1131–1135, 2022.

[28] 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.

[29] M. Hou, Z. Zhang, and G. Lu, "Multi-modal emotion recognition with self-guided modality calibration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4688–4692.

[30] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure." in *Interspeech*, 2021, pp. 1947–1951.

[31] L. Fu, X. Li, L. Zi, Z. Zhang, Y. Wu, X. He, and B. Zhou, "Incremental learning for end-to-end automatic speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 320–327.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[33] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[34] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[35] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[36] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.

[37] 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.