Article

Not peer-reviewed version

Speech Emotion Recognition Using Multi-Scale Global-Local Representation Learning with Feature Pyramid Network

Yuhua Wang , Jianxing Huang , Zhengdao Zhao , Haiyan Lan * , Xinjia Zhang

Posted Date: 14 October 2024

doi: 10.20944/preprints202410.1002.v1

Keywords: speech emotion recognition; multi-scale feature pyramid network; convolutional self-attention

 doi:10.20944/preprints202410.1002.v1

Article

Speech Emotion Recognition Using Multiscale Global-Local Representation Learning with Feature Pyramid Network

Yuhua Wang, Jianxing Huang, Zhengdao Zhao, Haiyan Lan * and Xinjia Zhang

College of Computer Science and Technology, Harbin Engineering University, Harbin, China
* Correspondence: lanhaiyan@hrbeu.edu.cn

Abstract: Speech emotion recognition (SER) is important in facilitating natural human-computer interactions. In speech sequence modeling, a vital challenge is to learn context-aware sentence expression and temporal dynamics of paralinguistic features to achieve unambiguous emotional semantic understanding. In previous studies, the SER method based on the single-scale cascade feature extraction module could not effectively preserve the temporal structure of speech signals in the deep layer, downgrading the sequence modeling performance. In this paper, we propose a novel multi-scale feature pyramid network to mitigate the above limitations. With the aid of the bi-directional feature fusion of the pyramid network, the emotional representation with adequate temporal semantics is obtained. Experiments on the IEMOCAP corpus demonstrate the effectiveness of the proposed methods and achieve competitive results under speaker-independent validation.

Keywords: speech emotion recognition; multi-scale feature pyramid network; convolutional self-attention

1. Introduction

Speech emotion recognition (SER) refers to extracting and analysing emotion-related features from speech signals, allowing the computer to understand the speaker's emotional expression [1]. As an essential part of affective computing, SER is the key to facilitating natural human-computer interaction (HCI) [2]. The HCI system with an affective computing module has been applied to many tasks, such as psychological assessment [3], mobile services [4], and safe driving [5]. Decades of research in SER have been devoted to modelling emotional representations from linguistic(e.g., lexical, syntactic, discourse and rhetorical features) and paralinguistic(e.g., supra-segmental phoneme and prosodic features) characteristics and developing the appropriate algorithms to implement robust and effective emotion recognition [6].

Recently, significant progress has been made in the field of SER using multi-scale convolutional neural networks (MSCNN). MSCNNs can extract multi-scale temporal and spatial features. Compared to single-scale network methods, MSCNN [10–14] effectively captures emotion-related features of variable lengths from speech inputs and is expected to further enhance semantic understanding and emotion recognition.

However, the MSCNN module in the previous research used a one-way transmission design. The temporal dynamic range that deep emotional features can retain depends on the size of the convolutional kernel. This unidirectional design can lead to the loss of temporal structure of speech in progressive resolution reduction. To enhance SER performance, it's crucial to learn both global-local semantic representation and temporal dynamics in speech. The main limitations are: 1) The high-level semantic features learned by the hierarchical MSCNN layers suffer from the loss of temporal structure of speech. 2) In context-aware semantic understanding, the MSCNN independently produces features that ignore the long-term relation between multi-scale sound units, and the correlations between multi-scale features should be explored.

Based on the limitations of the existing feature extraction networks in emotional feature extraction, we propose a novel framework learning context-aware representation using a multiscale feature pyramid network (MSFPN) for SER. We explore the bidirectional fusion of multi-scale semantic

 doi:10.20944/preprints202410.1002.v1

features using feature pyramid network and preserves the resolution for temporal dynamics learning and global-local semantic understanding. More specifically, we adopt parallel MSCNN groups for multi-scale feature learning, supplemented with a forward fusion mechanism to fuse these features into high-level semantic features. Moreover, to learn the local-global correlations in MSCNN, this paper improves and uses the convolutional self-attention (CSA) [15] layer to focus on the emotion-related periods in the local region effectively. MSFPN enhances the connections between adjacent elements and captures the interactions between features extracted by different attention heads. In the top-down pathway, features are rejoined into low-level acoustic features by the backward fusion, and BiLSTM is adopted to generate an utterance-level representation for emotion classification.

Our research contributions can be summarized as follows:

1. We have conducted a detailed exploration of the application of feature pyramids in speech emotion recognition for the first time. We enhanced the MSCNN by integrating the CSA module to better capture local emotional correlations.
2. We have improved the CSA by using a multi-scale convolutional module, avoiding the degradation problem of the convolutional attention network.
3. We designed a backward fusion approach that effectively captures features across different levels of detail, successfully preserving the importance of local dynamics and deep semantics in emotional representation.

The rest of the paper is organized as follows. We summarize related work in Section 2. In Section 3, we propose our framework in detail. Experiments are reported in Section 4. And we summarize this paper in Section 5.

2. Related Work

2.1. background of Speech Emotion Recognition

Emotions are psychological states triggered by neurophysiological changes, and they are related differently to thoughts, feelings, behavioral responses, and varying degrees of happiness or unhappiness. Currently, there is no scientific consensus on the definition of emotions, which often intertwine with feelings, temperament, personality, character, and creativity. Emotions consist primarily of subjective experiences, physiological responses, and behavioral reactions, playing a crucial role in building interpersonal relationships. They can be recognized through speech, facial expressions, and body language. Basic emotions include surprise, joy, disgust, sadness, anger, and fear, while complex emotions, such as contempt, amusement, and embarrassment, are blends of multiple feelings and are more challenging to identify.

Speech Emotion Recognition (SER) technology aims to identify human emotions through voice. Typically, people are not very accurate in recognizing others' emotions, making emotion recognition a burgeoning field of study where appropriate technology can enhance accuracy. The core of this process lies in identifying emotions from speech without considering cognitive content. SER mainly involves two processes: feature extraction and emotion classification, and it holds significant potential for applications in security, healthcare, entertainment, and education. The task of emotion recognition is complex due to the highly subjective nature of emotions, and there is yet to be a unified standard for classifying or measuring them.

As illustrated in Figure 1, an SER system consists of modules for speech signal input, preprocessing, feature extraction and selection, classification, and emotion recognition. The ability to extract robust emotional features from speech is crucial to the success of an SER system. Studies have shown that extracting features across multiple modal[35–41] and multiple scales can significantly improve the accuracy of emotion recognition.

 doi:10.20944/preprints202410.1002.v1

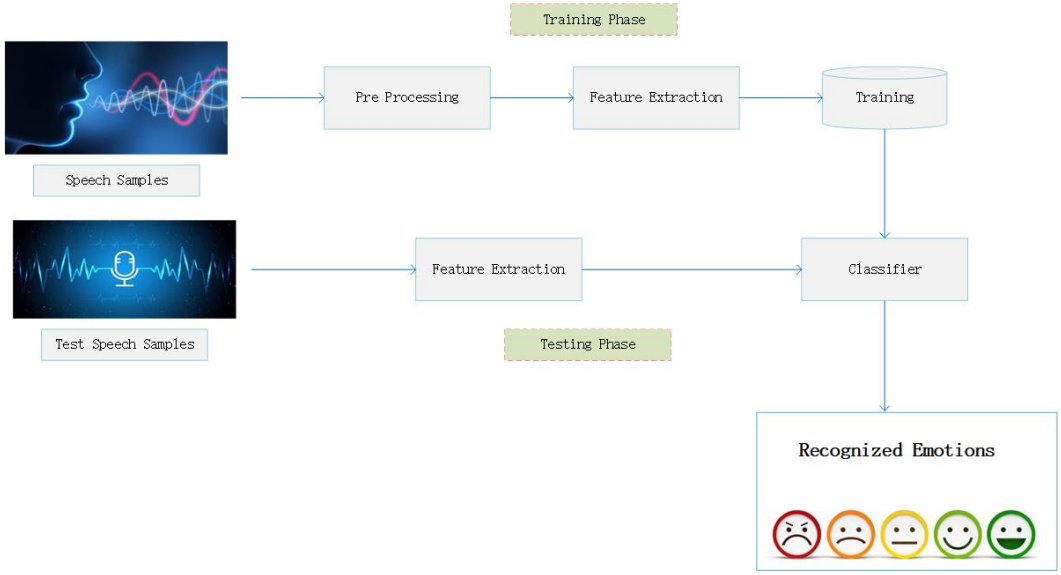


Figure 1. Functional diagram of SER system

2.2. Multi-Scale Network Model

A multi-scale network model is a deep learning architecture designed to capture and process features across different scales simultaneously, thereby enhancing overall model performance. Typically, the model consists of multiple branches, each handling input data at various scales. These features are then fused at specific stages, enabling the extraction of more robust and comprehensive representations.In speech emotion recognition (SER),where emotions are subjective and vary among individuals, many researchers focus on extracting emotion features from speech at multiple scales to enhance SER performance. Zhu et al. [12] utilized a global perceptual fusion method to extract features across various scales in speech emotion recognition. They created a special neural network that learns and combines these multi-scale features using a global perceptual fusion module. Peng et al. [10] proposed a framework using MSCNN with statistical pooling units to obtain both the audio and text hidden representations and employed an attention mechanism to improve performance further. Xie et al. [31] utilized multi-head attention to extract emotional features from both the temporal and frequency domains. They employed additive attention in the frequency domain, enhancing the capability to extract nonlinear features. In modelling temporal dependence, Chen et al. [14] employed MSCNN with a bidirectional long short-term memory network (BiL-STM) to model local-aware temporal dynamics for SER. Gan et al. Li et al. [22] utilized a multi-scale Transformer, incorporating multi-scale temporal feature operators, attention modules, and proportion mixers. This approach effectively extracts emotion features across different time scales, enhancing the accuracy of emotion recognition.

Nowadays, multi-scale feature extraction for emotion has become widely adopted. However, existing multi-scale networks often employ one-way transfers, which can compromise the temporal structure of speech. To tackle this, we've incorporated a bidirectional transfer design in our model, which better maintains the emotional features' temporal structure compared to other models.

2.3. feature Pyramid Network Model

The feature pyramid network [25] was initially designed to address multi-scale challenges in object detection. Due to its capability to effectively capture features at various granularities, it has become widely adopted in object detection tasks. Lately, researchers have been investigating how feature pyramid networks can be applied to sequential tasks like speech recognition and classification. Liu et al. [26] introduced a contextual pyramid generative adversarial network for speech enhancement. This design effectively captures speech details at various levels and removes audio noise in a structured way. Luo et al. [27] Explored sound event detection based on feature pyramid networks,

 doi:10.20944/preprints202410.1002.v1

and experiments demonstrated that the model utilizing feature pyramid networks surpassed LSTM in performance. Furthermore, researchers [28–30] have utilized feature pyramid networks to enhance speech recognition capabilities by combining multi-level features.

Previous studies have shown that feature pyramid networks can effectively capture multi-scale features in sequential tasks. Yet, there's been limited research on how these networks benefit extracting emotion features from speech. So, we delved deeper into using feature pyramid networks for speech emotion recognition. We enhanced the original feature pyramid network by introducing a more powerful convolutional attention network to extract stronger emotion features.

3. Methodology

In this section, we propose our framework in detail. The proposed framework, as illustrated in Figure 2, comprises two primary modules: the multi-scale feature pyramid model and the global-local representation learning model. The multi-scale feature pyramid model consists of three sets of convolutional blocks (Conv Blocks), each characterized by distinct kernel sizes. Specifically, the first set employs 3×3 kernels, the second set utilizes 5×5 kernels, and the third set incorporates 7×7 kernels. This varied kernel configuration facilitates the effective extraction of emotional features across different scales. Furthermore, the groups interact and fuse features through both forward and backward fusion mechanisms, enhancing the robustness of the extracted multi-scale emotional characteristics.The global-local representation learning model is composed of three bidirectional long short-term memory networks (BiLSTMs), which are specifically designed to capture global features from the outputs of the multi-scale feature pyramid model.

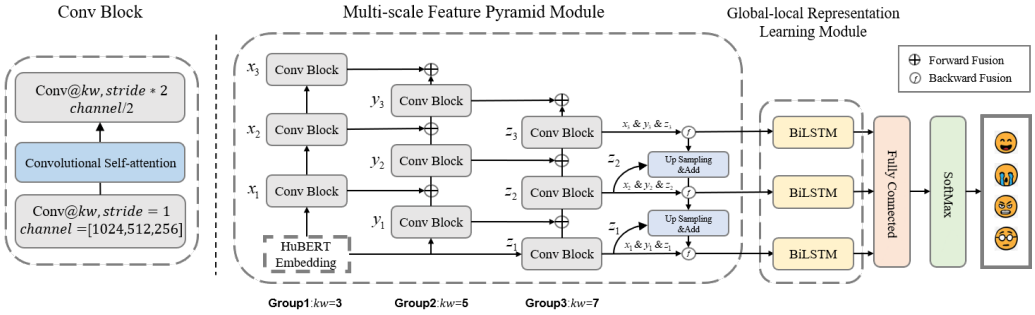


Figure 2. The overview of proposed multi-scale feature pyramid network.

The process commences with the input speech signal being processed through a pre-trained HuBERT model to generate HuBERT embeddings [16]. These embeddings are subsequently input into both the multi-scale feature pyramid model and the global-local representation learning model to extract multi-scale discourse-level features. Ultimately, the framework generates the final emotional output through two fully connected layers.

3.1. Multi-Scale Feature Pyramid Network

The overview architecture of the MSFPN is proposed as shown in Figure 2. The overall structure consists of two paths: one for extracting deep emotional features from bottom to top and another for merging multi-scale emotional features from top to bottom.

In the bottom-up pathway, as shown in Figure 3, the MSFPN includes a series of ConvBlock layers to learn and protrude the fine-grained emotion-related features in the different levels. Each ConvBlock contains two CNN layers for feature learning and is mixed with an improved convolutional self-attention (detailed in Section 3.2) layer for local correlation learning.

 doi:10.20944/preprints202410.1002.v1

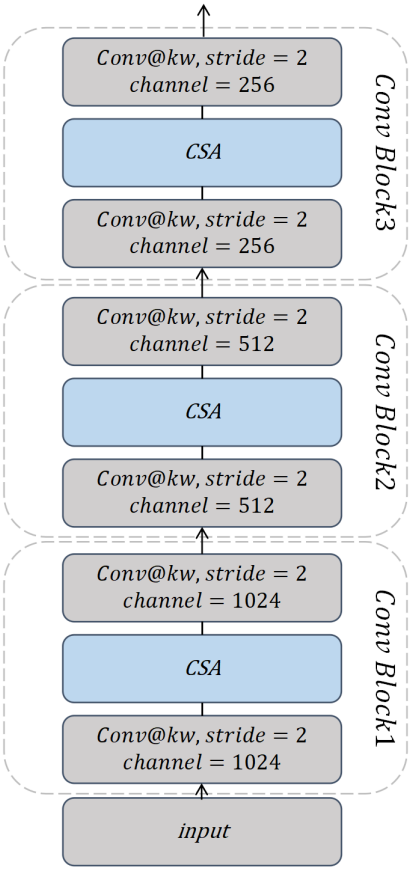


Figure 3. bottom-up pathway, where kw denotes different kernel widths, and CSA denotes convolutional self-attention

Specifically, there are three groups (X, Y, Z) of bottom-up pathways with different kernel widths (kw = 3, 5, 7), each extracting three levels of semantic features at different time scales. Given a speech dataset with n utterances, the characteristics of Group1 (x_1, x_2, x_3) are calculated using hierarchical ConvBlock layers shown in Figure 3. After each layer, the number of output channels is reduced by half. To be specific, the output channels for x_1 are 1024, for x_2 are 512, and for x_3 are 256. The deep features of Group2 (y_1, y_2, y_3) and Group3 (z_1, z_z, z_3) are computed in the same way. The difference is that y_i and z_i can receive the previous fine-grained acoustic emotional features from the above groups by a forward fusion to enhance the local semantic understanding of phrases. Typically, the forward fusion here is adding function.

In the top-down pathway, a backward fusion mechanism is introduced to combine semantic feature maps with high-resolution acoustic features. As shown in Figure 4,For deep features $F_i = [x_i, y_i, z_i]$, firstly, calculate the attention score $\alpha_i$for the i-th layer using Equation (1). Then, multiply α_i with features from various depth levels to extract the most pertinent emotion features for the i-th layer. Lastly, sum these products to derive the feature $G_i$for the i-th layer.

$$\alpha_i = softmax(W_2 \times Tanh(W_1 F_i^T)) \tag{1}$$

$$G_i = \sum_{i=1}^{3} \alpha_i \times F_i \tag{2}$$

where $G_i$is the multi-scale deep features of the i-th level in the MSFPN, W_1 and $W_2$are trainable parameters.

 doi:10.20944/preprints202410.1002.v1

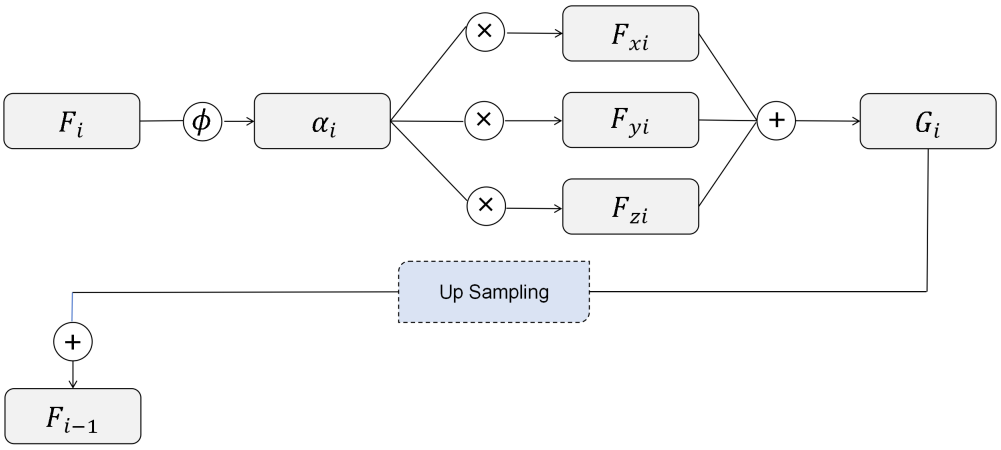
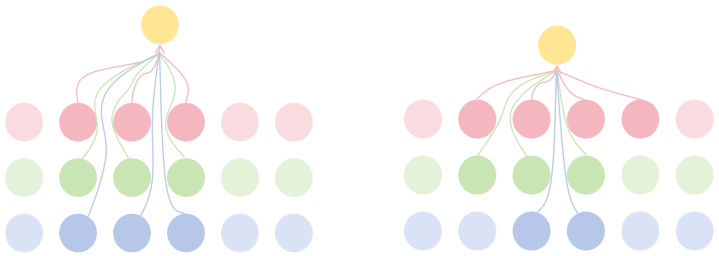


Figure 4. backward fusion structure,where ϕ represents the attention score calculation function as shown in Equation (1), and F_i denotes the feature of the i-th layer.

Then the multi-scale deep feature G is upsampled by a transposed convolution function, and the low-level features add the high-level semantic information. We acquire the deep features that contain adequate semantic information for understanding the emotional expression of each utterance and a high resolution with the sequential structure for temporal dependence learning.

3.2. Convolutional Self-Attention

A vanilla convolutional self-attention mechanism is proposed in [15] for learning short-distance dependencies. As illustrated in Figure 5(a), self-attention [17] disperses attention across all elements, which can lead to the neglect of relationships between adjacent elements and phrase patterns. In contrast, CSA effectively captures features between neighboring utterances, demonstrating a strong capacity for fine-grained feature extraction. However, the CSA with a fixed kernel size has limitations in fusing multi-scale features. When CSA projects multi-head spaces of the same dimension, the restricted regions of interest and the high number of heads can cause the regions projected by different heads to converge into similar feature spaces.



[a] [b]

Figure 5. convolutional self-attention(CSA) framework.(a) vanilla CSA ;(b)improved CSA

A viable solution is to employ multi-scale feature extraction to enhance feature space projection, allowing self-attention to compute temporal dependencies across different information regions while mitigating the overfitting problem associated with the multi-head mechanism in CSA. In this paper, we propose an improved convolutional self-attention layer that incorporates multi-scale feature extraction. As shown in Figure 5(b), we employ different kernel sizes across various attention heads, enabling the extraction of fine-grained features at multiple scales. Given an input sequence X, the CSA focuses

 doi:10.20944/preprints202410.1002.v1

on local periods for each query q_i and restricts its attention region to a local scope of fixed size $m+1$ ($m \in \{3,7,11,15\}$)), centered at position i. To maintain the same output dimension as the input for the CSA layer, we reduce the output dimension for each kernel width by a factor of four.

$$\hat{K_m} = \{k_{i-\frac{m}{2}}, \cdots, k_i, \cdots, k_{i+\frac{m}{2}}\}. \tag{3}$$

$$\hat{V_m} = \{v_{i-\frac{m}{2}}, \cdots, v_i, \cdots, v_{i+\frac{m}{2}}\}. \tag{4}$$

Each head's linear mapping query, key, and value with dimensions of dk, dk, and dv. In practice, self-attention computes the attention function on a set of queries simultaneously. The query, key, and value are packed together into a matrix Q, $\hat{K_m}$, and $\hat{V_m}$. The attention output is calculated as:

$$Attention(Q, \hat{K_m}, \hat{V_m}) = softmax(\frac{Q\hat{K_m}^T}{\sqrt{d_k}})\hat{V_m}. \tag{5}$$

$$Head_{i,m} = Attention(QW_i^Q, \hat{K_m}W_i^{\hat{K_m}}, \hat{V_m}W_i^{\hat{V_m}}) \tag{6}$$

$$MultiHead(Q, \hat{K_m}, \hat{V_m}) = Concat(head_{1,3}, \cdots, head_{l,m}) \tag{7}$$

where W_i^Q, $W_i^{\hat{K_m}}$, $W_i^{\hat{V_m}}$ are the weight matrices in multi-head attention with dimensions d_k/l, d_k/l , d_v/l, respectively.

3.3. Global-Local Representation Learning Module

In previous studies, the BiLSTM network has been proven effective in capturing the long temporal dynamics of deep features to aggregate global-local representations[14], but the progressive resolution reduction limits the sequence modeling performance[9]. In this paper, we explore the relative interaction between each emotional state in the progressive acoustic feature extraction. We model the multi-scale temporal dependence to generate the global-local representations for SER. In practice, the representations $\mathcal{R}$ is aggregated by BiLSTM $H = (h_1, h_2, \cdots, h_t)$ for the multi-scale deep features $\mathcal{G} = (g_1, g_2 \cdots, g_t)$.

$$\begin{aligned}
f_t &= \sigma(W_f[g_t, h_{t-1}] + b_f) \\
i_t &= \sigma(W_i[g_t, h_{t-1}] + b_i) \\
o_t &= \sigma(W_o[g_t, h_{t-1}] + b_o) \\
C_t &= f_t * C_{t-1} + i_t * tanh(W_c[g_t, h_{t-1}] + b_c) \\
h_t &= o_t * tanh(C_t) \\
\mathcal{R} &= h_t^1 \oplus h_t^2 \oplus h_t^3
\end{aligned} \tag{8}$$

Here, σ denotes the sigmoid activation function, while f, i, o, and C represent the vectors for the input gate, forget gate, output gate, and memory cell activation, respectively. The weight matrices and bias vectors for each gate are indicated by W and b. The last hidden output, h_t^i, serves as the utterance-level representation $\mathcal{R}$, which is subsequently input into fully connected layers for emotion inference.

4. Experiments

In this section, we evaluate the proposed framework in the IEMOCAP corpus. This paper compares the results with related state-of-the-art methods and deploys ablation studies to measure each component's contribution.

 doi:10.20944/preprints202410.1002.v1

4.1. Corpora Description

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus[18], commonly utilized for evaluation. IEMOCAP is an interactive emotional binary motion capture database developed by the SAIL Lab at the University of Southern California, containing a total of 12 hours of recordings. The data was recorded by ten professional actors (five males and five females) in a studio setting. Each recording is accompanied by discrete emotional labels and annotations for emotional dimensions. The IEMOCAP database consists of five sessions, each featuring dialogues between a female and a male actor, divided into two parts: improvised performances and scripted performances. The former involves spontaneous dialogues without predetermined content, while the latter follows a predefined script. This database encompasses multimodal information, including audio and text, making it suitable for various unimodal emotion recognition studies. To ensure balanced representation of audio samples across categories, this study combines the excitement emotion into the happiness category. Ultimately, the dataset comprises 5,531 audio samples, distributed as follows: 1,103 instances of anger, 1,084 instances of sadness, 1,708 instances of calmness, and 1,636 instances of happiness. Figure 6 illustrates the number of audio samples corresponding to each emotional label. As there is no predefined data split in IEMOCAP, we perform 10-fold cross-validation with a leave-one-person-out strategy to achieve comparable results.

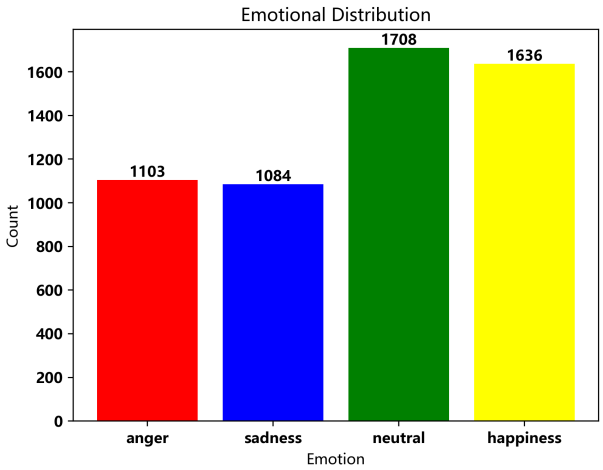


Figure 6. The number of audio samples corresponding to each emotional label in IEMOCAP

4.2. Implementation Details

The proposed framework was implemented in PyTorch. For training, we used the Adam optimizer with a learning rate of 1e-5, a batch size of 32, and applied early stopping [19]. To handle data imbalance, we evaluated performance using both weighted accuracy (WA) and unweighted accuracy (UA). WA measures the overall classification accuracy by dividing the total number of correctly predicted samples by the total number of samples, while UA calculates the average accuracy for each emotion category, providing insight into the model's performance across different emotions.

$$\begin{aligned} WA &= \frac{\sum_{i=1}^{K} N_i * Accuracy_i}{\sum_{i=1}^{K} N_i}. \\ UA &= \frac{1}{K} \sum_{i=1}^{K} Accuracy_i. \end{aligned} \tag{9}$$

where the K denotes the number of emotion categories and i represents the i-th emotion category. N_i denotes the data quantity of the i-th emotion category.

 doi:10.20944/preprints202410.1002.v1

4.3. Experiment Results and Discussion

To fairly compare the performance of the proposed framework, we implement the end-to-end method (E2ESA without multi-task learning), resolution maintained method (DRN), and multi-scale feature representation method (GLAM without data augmentation) for evaluation. The experimental results are shown in Table 1. In the MSCNN-based method, GLAM's multiple feature representations are conducted on the high-level semantic features, which are limited by the fine-grained temporal dynamics and local representation learning. Compared to the end-to-end method, E2ESA uses single-scale correlation modelling and limits global-local representation learning for context-aware emotion classification. The DRN method appropriately preserves the resolution of deep features, which avoids the loss of temporal structure of speech in hierarchical CNN. However, directly applying the MSCNN module in the DRN leads to gridding effect when learning high-level temporal dynamics upon the mediate hidden features. In this paper, the *MSFPN* mainly benefits from the high semantic and full resolution feature map for global-local representation learning and achieves the highest results with 3.69% UA (GLAM),3.39% UA (Xie[31]), 2.53%UA (E2ESA), and 1.8%UA (DRN) improvements, respectively.

Table 1. Comparisons of UA and WA with state-of-the-art methods on IEMOCAP. The best results are highlighted in bold.

Model	UA	WA
GLAM[12]	69.70%	68.75%
E2ESA[31]	70.86%	69.25%
Xie[20]	70.0%	68.8%
DRN[9]	71.59%	70.23%
MSFPN	**73.39%**	**71.79%**

In addition, the feature space of *MSFPN* is well compacted and assembled. The t-SNE visualization is depicted in Figure 7. This figure shows that our method can produce differentiated emotional features. Unlike DRN, neutral emotional features are more condensed and refined and less mixed in other emotional spaces.

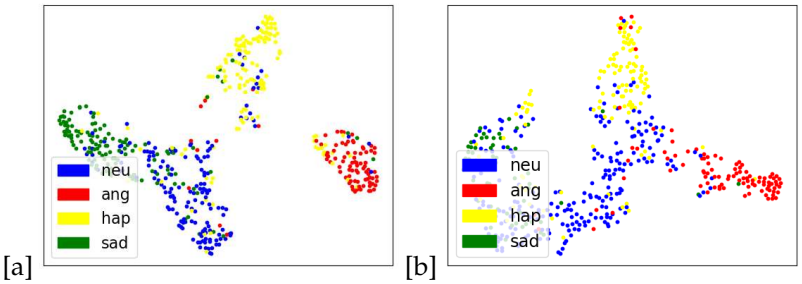


Figure 7. The t-SNE visualization of the proposed framework.(a) MSFPN;(b)DRN

4.4. Ablation Study

To further measure the contributions of each component of the proposed model, the results of ablation experiments are shown in Table 2.

 doi:10.20944/preprints202410.1002.v1

Table 2. Performance of ablation studies on IEMOCAP dataset. 'w/o' denotes the vanilla *MSFPN* framework without certain component.

Method	UA	WA
w/o CSA	66.89%	65.35%
w/o MSCNN	70.95%	69.05%
w/o forward fusion	71.85%	70.09%
w/o backward fusion	72.72%	71.26%
MSFPN	**73.39%**	**71.79%**

From the results in Table 2, there is a clear measurement of the contributions of each component. The forward fusion provides the fusion of multi-scale deep features for robust semantic feature leaning in progressive downsampling. The backward fusion mainly focuses on the salient emotional periods in multi-scale deep features and aggregates them into a high-resolution feature map. The CSA in MSCNN is important for local correlation learning, which effectively enhances the expression of emotion in semantically strong features. The core component in our proposed framework is the MSCNN, which extracts multiple deep features and enables the fusion mechanism to appropriately capture the emotional-related features.

5. Conclusion

In this paper, we propose a multi-scale Feature Pyramid Network for context-aware speech emotion recognition. The MSCNN-based feature extraction module with an improved CSA layer is useful for capturing global-local correlations in the speech sequence. With bottom-up and top-down connection, semantically strong features are effectively aggregated with high-resolution acoustic features, which contains adequate emotional characteristics and retain temporal structure for global-local representation learning. The experimental results on IEMOCAP demonstrate our framework's effectiveness and significantly improved over the state-of-the-art approaches.

References

1. Korsmeyer, C. Rosalind W. Picard, affective computing. In Minds and Machines, 1999, vol. 9, no. 3, pp. 443–447.
2. Schuller, B. W. Speech emotion recognition two decades in a nutshell, benchmarks, and ongoing trends. In Communications of the ACM, 2018, vol. 61, no. 5, pp. 90–99.
3. Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L., & Allen, N. B. Detection of clinical depression in adolescents' speech during family interactions. In IEEE Transactions on Biomedical Engineering, 2011, vol. 58, no. 3, pp. 574–586.
4. Yoon, W.-J., Cho, Y.-H., & Park, K.-S. A study of speech emotion recognition and its application to mobile services. In Ubiquitous Intelligence and Computing, 2007, pp. 758–766, Springer Berlin Heidelberg.
5. Tawari, A., & Trivedi, M. Speech based emotion classification framework for driver assistance system. In 2010 IEEE Intelligent Vehicles Symposium, 2010, pp. 174–178.
6. Ma, H., & Yarosh, S. A review of affective computing research based on function-component-representation framework. In IEEE Transactions on Affective Computing, 2021, pp. 1–1.
7. Basu, S., Chakraborty, J., & Aftabuddin, M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017, pp. 333–336.
8. Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., & Schmauch, B. CNN+LSTM architecture for speech emotion recognition with data augmentation. In Workshop on Speech, Music and Mind (SMM 2018), Sep 2018.
9. Li, R., Wu, Z., Jia, J., Zhao, S., & Meng, H. Dilated residual network with multi-head self-attention for speech emotion recognition. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6675–6679.

10. Peng, Z., Lu, Y., Pan, S., & Liu, Y. Efficient speech emotion recognition using multi-scale CNN and attention. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 3020–3024.
11. Liu, J., Liu, Z., Wang, L., Guo, L., & Dang, J. Speech emotion recognition with local-global aware deep representation learning. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7174–7178.
12. Zhu, W., & Li, X. Speech emotion recognition with global-aware fusion on multi-scale feature representation. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6437–6441.
13. Xu, M., Zhang, F., Cui, X., & Zhang, W. Speech emotion recognition with multiscale area attention and data augmentation. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6319–6323.
14. Chen, M., & Zhao, X. A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. In Proc. Interspeech 2020, 2020, pp. 374–378.
15. Yang, B., Wang, L., Wong, D. F., Chao, L. S., & Tu, Z. Convolutional self-attention networks. In CoRR, 2019, vol. abs/1904.03107.
16. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. Self-supervised speech representation learning by masked prediction of hidden units. In IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, vol. 29, pp. 3451–3460.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. Attention is all you need. In arXiv, 2017.
18. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. IEMOCAP: Interactive emotional dyadic motion capture database. In Language Resources and Evaluation, 2008, vol. 42, no. 4, pp. 335–359.
19. Prechelt, L. Early stopping—but when? In Neural Networks: Tricks of the Trade, 1998, pp. 55–69, Springer-Verlag.
20. Li, Y., Zhao, T., & Kawahara, T. Improved end-to-end speech emotion recognition using self-attention mechanism and multitask learning. In Proc. Interspeech 2019, 2019, pp. 2803–2807.
21. Gan, C., Wang, K., Zhu, Q., Xiang, Y., Jain, D. K., & García, S. Speech emotion recognition via multiple fusion under spatial–temporal parallel network. In Neurocomputing, 2023, 555.
22. Li, Z., Xing, X., Fang, Y., Zhang, W., Fan, H., & Xu, X. Multi-scale temporal transformer for speech emotion recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-August, pp. 3652–3656.
23. Xie, Y., Liang, R., Liang, Z., Zhao, X., & Zeng, W. Speech emotion recognition using multihead attention in both time and feature dimensions. In IEICE Transactions on Information and Systems, 2023, E106.D(5), pp. 1098–1101.
24. Yu, L., Xu, F., Qu, Y., et al. Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion. In Applied Acoustics, 2024, 216: 109752.
25. Lin, T. Y., Dollár, P., Girshick, R., et al. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
26. Liu, G., Gong, K., Liang, X., et al. CP-GAN: Context pyramid generative adversarial network for speech enhancement. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6624–6628.
27. Luo, S., Feng, Y., Liu, Z. J., et al. High precision sound event detection based on transfer learning using transposed convolutions and feature pyramid network. In 2023 IEEE International Conference on Consumer Electronics (ICCE), 2023, pp. 1–6.
28. Basbug, A. M., & Sert, M. Acoustic scene classification using spatial pyramid pooling with convolutional neural networks. In 2019 IEEE 13th International Conference on Semantic Computing (ICSC), 2019, pp. 128–131.
29. Gupta, S., Karanath, A., Mahrifa, K., et al. Segment-level probabilistic sequence kernel and segment-level pyramid match kernel based extreme learning machine for classification of varying length patterns of speech. In International Journal of Speech Technology, 2019, 22: 231–249.

30. Ren, Y., Peng, H., Li, L., et al. A voice spoofing detection framework for IoT systems with feature pyramid and online knowledge distillation. In Journal of Systems Architecture, 2023, 143: 102981.
31. Xie, Y., Liang, R., Liang, Z., et al. Speech emotion recognition using multihead attention in both time and feature dimensions. In IEICE Transactions on Information and Systems, 2023, 106(5): 1098–1101.
32. Manelis, A.; Miceli, R.; Satz, S.; Suss, S.J.; Hu, H.; Versace, A. The Development of Ambiguity Processing Is Explained by an Inverted U-Shaped Curve. Behav. Sci. 2024, 14, 826. https://doi.org/10.3390/bs14090826
33. Arslan, E.E.; Akşahin, M.F.; Yilmaz, M.; Ilgın, H.E. Towards Emotionally Intelligent Virtual Environments: Classifying Emotions through a Biosignal-Based Approach. Appl. Sci. 2024, 14, 8769.
34. https://doi.org/10.3390/app14198769 Sun, L.; Yang, H.; Li, B. Multimodal Dataset Construction and Validation for Driving-Related Anger: A Wearable Physiological Conduction and Vehicle Driving Data Approach. Electronics 2024, 13, 3904. https://doi.org/10.3390/electronics13193904
35. Lee, J.-H.; Kim, J.-Y.; Kim, H.-G. Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning. Bioengineering 2024, 11, 997. https://doi.org/10.3390/bioengineering11100997
36. Liu, G.; Hu, P.; Zhong, H.; Yang, Y.; Sun, J.; Ji, Y.; Zou, J.; Zhu, H.; Hu, S. Effects of the Acoustic-Visual Indoor Environment on Relieving Mental Stress Based on Facial Electromyography and Micro-Expression Recognition. Buildings 2024, 14, 3122. https://doi.org/10.3390/buildings14103122
37. Das, A.; Sarma, M.S.; Hoque, M.M.; Siddique, N.; Dewan, M.A.A. AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition. Sensors 2024, 24, 5862. https://doi.org/10.3390/s24185862
38. Udahemuka, G.; Djouani, K.; Kurien, A.M. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. Appl. Sci. 2024, 14, 8071. https://doi.org/10.3390/app14178071
39. Zhang, S., Yang, Y., Chen, C., et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. In Expert Systems with Applications, 2024, 237: 121692.
40. Wang, Y., Li, Y., Cui, Z. Incomplete multimodality-diffused emotion recognition. In Advances in Neural Information Processing Systems, 2024, 36.
41. Meng, T., Shou, Y., Ai, W., et al. Deep imbalanced learning for multimodal emotion recognition in conversations. In IEEE Transactions on Artificial Intelligence, 2024.