Sensors and Materials, Vol. 38, No. 5 (2026) 2477–2490
MYU Tokyo

S & M 4448

Multiscale Network Leveraging Wavelet Scattering Feature Maps for Emotion Recognition from Acoustic Sensor Data

Na Ying,[1*] Mengfan Yu,[1] Shunpeng Wu,[1] Xinyu Lin,[1] Du Jiang,[2] and Yinfeng Fang[1]

[1]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2]Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China

(Received January 9, 2026; accepted April 9, 2026)

Keywords: human–robot interaction, emotion recognition, multiscale neural network, high-resolution networks, wavelet scattering feature maps

Reliable emotion recognition from acoustic sensor signals is a pivotal yet challenging component for achieving natural and empathetic human–robot interaction. We propose a dedicated, deployable audio processing module. The core of this module is a novel integration of wavelet scattering transform (WST) as a robust preprocessing front-end with a multiscale neural network. Specifically, WST standardizes raw, sensor-acquired audio signals into translation-invariant and deformation-stable feature maps, effectively mitigating input variability. Building upon this, we construct a wavelet scattering feature map–multiscale network (WSM-MSN) that synergistically combines hierarchical convolutional neural network (CNN) branches for extracting fine-grained local affective features with a high-resolution network (HRNet) branch to capture and fuse multiscale contextual dependences, thereby significantly enhancing recognition precision. Extensive evaluations on four datasets of affective acoustic signals (EMODB, RAVDESS, IEMOCAP, and eNTERFACE'05) demonstrate the module's superiority. This achieves consistent unweighted average recall (UAR) improvements of 1.62, 5.28, 4.44, and 3.37%, respectively, over traditional scattering methods and surpasses other comparative algorithms.

1. Introduction

The performance of microphone-based acoustic sensors in intelligent robotic systems, such as service or collaborative robots, is often compromised by environmental noise and signal variability, posing a fundamental challenge to reliable multimodal interaction. To address this sensor-level robustness issue in human–robot interaction (HRI), in this study, we propose a signal processing module. The module is designed to enhance the reliability of emotion inference from raw, sensor-acquired speech signals, which is a critical capability for empathetic HRI.

*Corresponding author: e-mail: yingna@hdu.edu.cn
https://doi.org/10.18494/SAM6134

ISSN 0914-4935
https://myukk.org/

Conventional approaches to analyzing such signals, including those based on mel-frequency cepstral coefficients (MFCCs), often struggle with noise and lack stability across diverse recording conditions. This work introduces a solution centered on wavelet scattering transform (WST), [1] reconceptualizing it not merely as an audio feature extractor but as a standardized preprocessing front-end for acoustic sensors. WST generates translation-invariant and deformation-stable feature maps directly from raw sensor signals, effectively mitigating intrinsic noise and variability.

Building upon this robust preprocessing foundation, we construct a wavelet scattering feature map–multiscale network (WSM-MSN) as the core computational unit. This architecture synergistically integrates a hierarchical convolutional neural network (CNN) for extracting fine-grained local features with a high-resolution network (HRNet) branch that maintains multiresolution representations to capture broader contextual dependences. This multiscale fusion is specifically designed to consolidate affective information from sensor signals with enhanced precision. The following core innovations are proposed in this research:

- a sensor-oriented preprocessing framework that employs WST to convert unstable, raw acoustic sensor outputs into robust, standardized feature maps;
- an MSN architecture optimized for processing the aforementioned feature maps, which efficiently fuses local and contextual information to improve the robustness of affective state recognition from sensor data;
- extensive evaluations on four benchmark datasets (EMODB,[2] RAVDESS,[3] IEMOCAP,[4] and eNTERFACE'05[5]) to underpin the module's superior generalization across varied acoustic environments.

The paper is organized as follows. In Sect. 2, the related work is reviewed. The architecture and training pipeline of the proposed WSM-MSN are detailed in Sect. 3. In Sect. 4, we evaluate the model using four benchmark datasets and compare its performance with those of state-of-the-art methods. Section 5 concludes the paper, and we discuss limitations and outline future research directions.

2. Related Work

2.1 Existing models

The pursuit of reliable emotion recognition from acoustic sensor data, particularly in noisy robotic environments, has intensified with advances in deep learning and the availability of diverse datasets.

Wang *et al.* investigated the problem of spectrogram-based speech emotion recognition for multiple corpora through a transfer learning approach using the residual network (ResNet) architecture and the residual adapter technique.[6] Subsequently, Falahzadeh *et al.* suggested a novel approach in which the initial audio waveform is transformed into a three-dimensional (3D) tensor representation by reconstruction phase space (RPS) and input into a 3D CNN, which is proven to capture the emotional features effectively.[7] Singh *et al.* compared and analyzed the performance of short-term MFSCs and constant-Q transform (CQT) across four benchmark

datasets through a deep neural network (DNN), and found that CQT has an advantage in low-frequency resolution, which in turn improves the emotion recognition performance.[8] Falahzadeh *et al.* further used a transfer learning strategy to input the 3D projection of the original waveform RPS into a visual geometry group 16 model and tuned the hyper-parameters using the Grey Wolf optimization algorithm to significantly enhance accuracy.[9] Gerczuk *et al.* used a parallel CNN to learn temporal and spectral information from MFCCs to further simplify the model complexity.[10] Mohan *et al.* developed an affective computing framework utilizing a two-dimensional CNN integrated with extreme gradient enhancement, which demonstrated good sentiment recognition accuracy on the RAVDESS dataset.[11] Flower and Jaya proposed a two-channel feature fusion network based on MFCCs and mel-frequency magnitude coefficients of a one-dimensional CNN to improve the effectiveness in speaker-independent emotion recognition.[12] Li *et al.,* on the basis of the dual stream network of MFCCs, chose Transformer architecture to capture long-range contextual dependences within acoustic sequences.[13] Dabbabi and Mars proposed a self-supervised framework integrating DistilHuBERT with multimodal audio-visual representations. MFCC features are extracted to encode speech characteristics for DistilHuBERT processing, while shared hidden representations across modalities enhance feature learning. This approach improves accuracy in both offline and real-time evaluations.[14] Chowdhury *et al.* proposed a lightweight deep neural integration model and conducted experiments using handcrafted features such as MFCC features to validate the efficacy of the integration model in automatically extracting handcrafted features.[15]

Beyond conventional input representations, WST has emerged as a promising feature extraction technique in voice processing domains. Singh *et al.* first applied WST to speech emotion recognition. Experiments on standard datasets indicate that WST features surpass MFCC-based features in characterizing emotional information.[16] Kek *et al.* used sub-spectral mixing and temporal mixing based on multi-timescale wavelet scattering features to classify sound scenes.[17] Sun *et al.* integrated wavelet scattering coefficients with psychoacoustic quality metrics to demonstrate the complementary benefits of these feature modalities for emotion recognition using Support Vector Machine (SVM).[18] In addition, Yu and Li constructed a wavelet scattering coefficient map that was subsequently processed into feature vectors via grey gradient covariance matrix computation and proved the feasibility of the proposed feature extraction method by comparing the features of different voices.[19] Cheng *et al.* proposed a high-resolution distance image target recognition algorithm based on WST to verify the effectiveness of WST in the case of small samples.[20]

2.2 WST

WST implements a hierarchical feature extraction process through three fundamental operations: convolution with wavelet filters, nonlinear transformation via complex modulus, and local averaging through low-pass filtering. This cascaded architecture systematically decomposes speech signals into multiscale representations while maintaining stability to signal deformations. Mathematically, the scattering coefficients are computed as described below.

In the first step, a low-pass filter is used to convolve the input signal $x(t)$ to obtain the zero-order scattering coefficient as

$$S_0 x(t) = x(t) * \phi_T(t), \tag{1}$$

where $x(t)$ is the input signal, $S_0x(t)$ is the low-pass filter, T is the time window length, and the symbol * denotes the convolution operation.

In the second step, $x(t)$ is convolved with the first-order wavelet filter bank and complex modes are solved to generate the first-order scale-map coefficients as

$$x_1(t, \lambda_1) = \left| x(t) * \psi_{\lambda_1}(t) \right|, \tag{2}$$

where $|\bullet|$ is the mode-taking operation; $\psi_{\lambda_1}(t)$ is a first-order wavelet filter bank that produces a cluster of bandpass filters with $\lambda_1 = 2^{j_1 + k/Q}$ as the center frequency by scaling the parent wavelet $\psi_\lambda(t)$, where Q is the quality factor (the number of wavelet filters in each bank per octave); $j_1 \in Z$ and $k \in (1, 2, \cdots, Q)$ denote the octave and color scale, respectively.

In the third step, to ensure the time-shift invariance of the signal, the first-order scattering coefficients are obtained by averaging the first-order scale-map coefficients using a low-pass filter as

$$S_1 x(t, \lambda_1) = x_1(t, \lambda_1) * \phi_T(t) = \left| x(t) * \psi_{\lambda_1}(t) \right| * \phi_T(t). \tag{3}$$

Similarly, the second-order scale-map coefficients are obtained by convolving $x_1(t, \lambda_1)$ with a second-order wavelet filter bank for complex modes, while the second-order scattering coefficients are obtained by averaging the second-order scale-map coefficients using a low-pass filter. The formula is

$$S_2 x(t, \lambda_2) = \left| \left| x(t) * \psi_{\lambda_1}(t) \right| * \psi_{\lambda_2}(t) \right| * \phi_T(t). \tag{4}$$

Repeating this algorithmic sequence, the t-th-order scattering coefficient is obtained when the iteration order is r and $r > 1$.

$$S_r x(t, \lambda_1, \cdots, \lambda_r) = \left| \left| \left| x(t) * \psi_{\lambda_1}(t) \right| * \cdots \right| * \psi_{\lambda_r}(t) \right| * \phi_T(t) \tag{5}$$

The multiscale feature extraction framework utilizes a hierarchical WST to capture signal characteristics across abstraction levels. This cascaded decomposition enables comprehensive signal representation, offering greater computational efficiency and feature interpretability than conventional DNNs while retaining competitive discriminative power for emotion recognition. Its inherent multi-resolution analysis makes it well suited to applications for balancing performance and interpretability.

3. WSM-MSN Algorithm

3.1 Data preprocessing

To adapt to the characteristics of microphone sensor signal acquisition and ensure the effectiveness of subsequent feature extraction and model training, this module employs the method described in literature during the preprocessing stage,[21] which is elaborated in detail as follows.

First, to ensure consistency in signal frequency representation, all speech samples captured by the microphone are resampled to 16 kHz. Second, to maintain uniformity in the temporal dimension of the input data, each recording is truncated to its initial 3-s-long segment. For samples with a duration of less than 3 s, zero padding is applied at the end of the signal until the total number of sampling points reaches the equivalent of 3 s. Finally, waveform normalization is performed on the processed signals. Specifically, the amplitude of each sample is scaled to a unified range by dividing by the maximum absolute value of the sample, thereby normalizing the signal amplitude range to the interval of [−1, 1].

3.2 WSM

After preprocessing, speech signals are transformed via a second-order wavelet scattering network to generate discriminative input features.

Assuming that the input speech signal is represented as $x(t)$, the zero-order wavelet scattering coefficient S_0 is computed using Eq. (1), the first-order wavelet scattering coefficient S_1 is obtained from Eq. (3) , and the second-order wavelet scattering coefficient S_2 is generated by Eq. (4). The obtained wavelet scattering coefficients form a second-order scattering feature matrix SC expressed as

$$SC = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \end{bmatrix}, \tag{6}$$

where the dimensions of SC are $N_q \times N_s$, N_q denotes the number of scattering paths, and N_s denotes the number of scale dimensions.

On the basis of the experimental setup and Ref. 13, WST is configured as follows. The parameter Q_1 in Eq. (2) controls the number of wavelet filters per octave in the first-order scattering transform and is set to 4. This value balances frequency resolution and computational efficiency, effectively capturing the local spectral structure of speech signals while avoiding excessive feature dimensionality. The parameter Q_2 in Eq. (4) controls the number of wavelet filters per octave in the second-order scattering transform and is set to 1. This sparse configuration is aimed at extracting higher-order time–frequency modulation information while effectively controlling feature dimensionality growth and avoiding redundancy. The invariance scale of T is set to 0.75 s. This value aligns with the critical time scale of emotional expression,

enhancing feature robustness to temporal perturbations while preserving the fine-grained temporal dynamics essential for emotion discrimination. To verify the rationality of the above parameter settings, we perform comparative experiments to evaluate their effects on feature extraction and classification performance, and present the results in Fig. 1.

The stability results in Fig. 1 are used to first analyze the rationality of the selected WST parameters. The stability metric is computed by adding Gaussian noise to the signal and measuring the energy change of scattering coefficients, with values closer to 1 indicating greater robustness. When $Q_1 = 4$, the stability reaches 0.8697, outperforming other values, confirming its optimal balance between frequency resolution and efficiency. When $Q_2 = 1$, the stability is 0.8068, which is slightly higher than those when $Q_2 = 2$ and $Q_2 = 4$, showing that sparse Q_2 captures modulation information without redundancy. When $T = 0.75$ s, the stability is 0.8072, which is close to that at the optimal $T = 0.5$ s and higher than that at $T = 1.0$ s, confirming alignment with the temporal scale of emotion expression. These results validate the parameter selection. Further analysis of noise robustness shows that the target parameters maintain high stability under noise, demonstrating strong anti-interference capability.

Using the above parameters, this matrix is then transposed and reshaped into a four-dimensional array of shape [K, 200, 24, 1], where K is the number of samples, and the 2D slices of size are extracted to generate WSMs by visualization.

Figure 2 shows the WSMs corresponding to eight distinct emotional states: anger, calm, disgust, fear, happy, neutral, sad, and surprise. The color indicates energy intensity levels, with darker hues corresponding to higher energy concentrations.

3.3 WSM-MSN network model

In general, the CNN structure can discriminate different emotions, but the excessive layering of convolutional operations frequently results in parameter redundancy. Therefore, inspired by

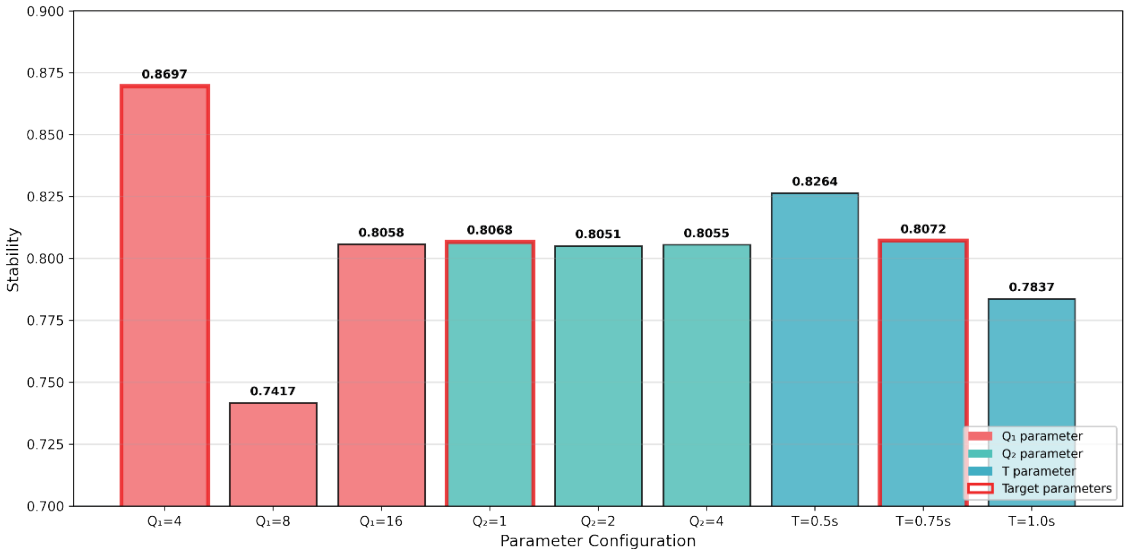

Fig. 1. (Color online) Stability comparison results.

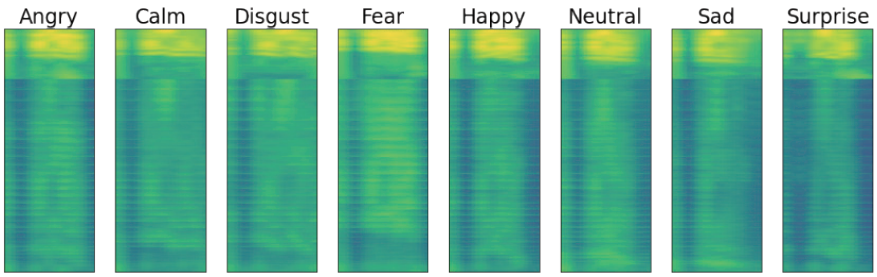

Fig. 2. (Color online) WSMs for different sentiment categories in the RAVDESS dataset.

HRNet as well as the literature,[22] to better recognize emotion features, we develop a novel affective computing framework utilizing WSM-MSN, which introduces a multiscale fusion method on top of the CNN to extend the network's local feature capturing capability. Figure 3 illustrates the complete workflow of the proposed method.

Figure 3(a) depicts the preprocessing stage that generates WSMs. Figure 3(b) presents the CNN branch, where the feature map passes through three convolutional layers (16, 32, 64 channels) to produce a 64-dimensional vector. Figure 3(c) shows the high-resolution branch, employing up-/down-sampling for feature reconstruction and fusion, also outputting a 64-dimensional vector. Figure 3(d) outlines the classification step: features from both branches are concatenated and fed into a dense layer for final emotion classification.

3.3.1 CNN branch

The proposed architecture employs a three-tier CNN branch for hierarchical feature extraction. Each stage consists of a 2D convolutional (Conv2D) layer (3 × 3 kernels) followed by average pooling (2 × 2) and batch normalization (BN), with channel sizes progressively increasing from 16 to 32 to 64 across stages. After the convolutional blocks, spatial features are flattened and projected into 64 dimensions.

3.3.2 High-resolution branch

The architecture features a three-stage, multiscale processing design. At each stage, input features are processed in parallel: a high-resolution path preserves details via Conv2D, while a low-resolution path captures context through down-sampling. The resulting features are concatenated and adjusted via 1 × 1 convolutions with progressively increasing size of channels (16, 32, 64). In the final stage, the features are condensed into a 64-dimensional representation.

4. Algorithm Validation Experiments

To systematically evaluate the robustness and generalization capability of the algorithm in recognizing emotions when confronted with different acoustic signals, this study designs the following experiments to verify its effectiveness.

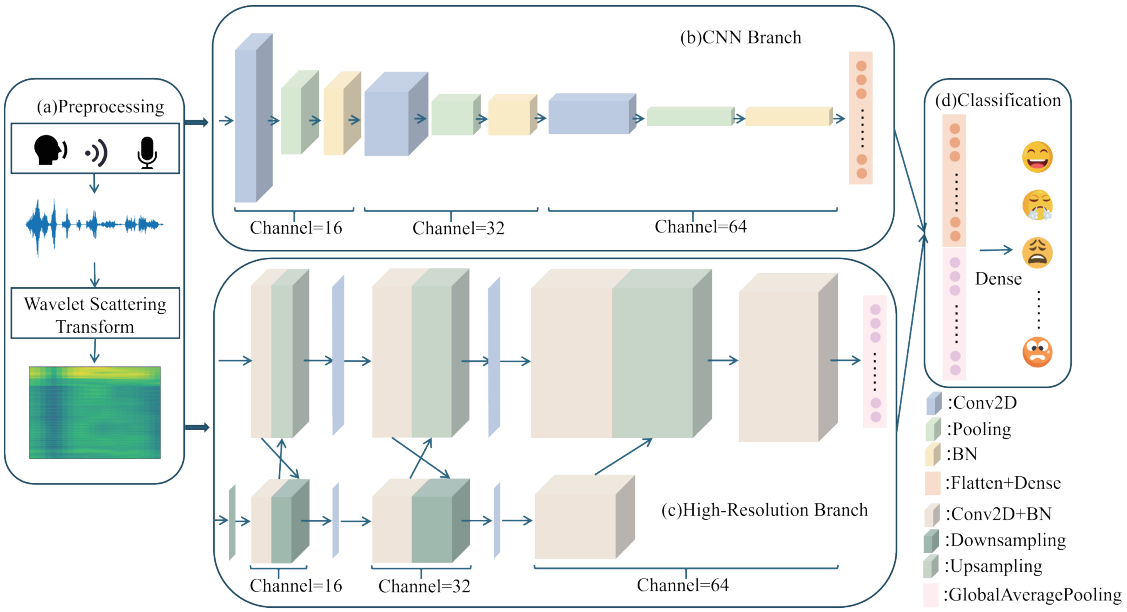

Fig. 3. (Color online) Flowchart of WSM-MSN algorithm. Here, "Channel" refers to the number of channels for each operation.

4.1 Databases

The experimental validation in this study is conducted on four publicly available datasets of affective acoustic signals. These datasets consist of raw audio recordings captured by microphone sensors, serving as standardized benchmarks to evaluate the robustness and generalization of the proposed processing module under various acoustic conditions. The specifications of each dataset are summarized in Table 1.

4.2 Experimental parameters

All experiments are implemented using the TensorFlow toolbox and conducted on a hardware platform equipped with an Intel(R) Xeon(R) CPU @ 2.30GHz, 12GB of RAM, and an NVIDIA Tesla T4 GPU. For the experimental setup, both the CNN and HRNet branches were configured with a batch size of 30, trained for 50 epochs using a learning rate of 0.001, and optimized with the Adam algorithm.

4.3 Evaluation metrics

Two key metrics are used to evaluate the model's performance: weighted average recall (WAR) and unweighted average recall (UAR).

$$WAR = \frac{\sum_{i=1}^{C} A_{ii}}{\sum_{i=1}^{C}\sum_{j=1}^{C} A_{ij}} \tag{7}$$

Table 1
Emotion distribution in dataset.

Dataset	Emotions	Modality
EMODB	Anger, neutral, fear, sadness, boredom, disgust, happiness	Audio
RAVDESS	Anger, happiness, disgust, sadness, boredom, fear, calmness, surprise	Audio
IEMOCAP	Happiness, neutral, sadness, anger	Audio, video
eNTERFACE'05	Surprise, fear, anger, disgust, sadness, happiness	Audio, video

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{A_{ii}}{\sum_{j=1}^{C}A_{ij}} \tag{8}$$

Here, A is the confusion matrix, C is the number of sentiment categories, and A_{ij} refers to the number of samples of category i that are classified as category j.

For WAR, the numerator corresponds to the sum of the diagonal elements of the confusion matrix, i.e., the total number of correctly classified samples; the denominator is the sum of all elements in the confusion matrix, i.e., the total number of samples in the dataset. WAR essentially measures the overall classification accuracy of the model across all categories.

For UAR, the calculation process is as follows: first, the recall for each class is computed separately, i.e., the ratio of the number of correctly classified samples of that class to the total number of actual samples in that class; then, the recall values of all classes are summed and divided by the total number of classes, yielding the arithmetic mean of the class-wise recall values. UAR provides a more equitable evaluation of the model's recognition capability across different classes.

4.4 Comparative experiments of related algorithms

To further validate the generalization of the algorithm, the proposed approach's efficacy is benchmarked against the baseline algorithm based on MFCCs and ScatNet. [16] The obtained experimental outcomes are presented in Table 2.

As shown by the results, the ScatNet method, which employs WST as its front-end feature extractor, consistently outperforms conventional MFCC-based methods across all four datasets, with average improvements of 11.12% in WAR and 11.19% in UAR. This clearly validates the effectiveness of WST preprocessing in capturing discriminative and robust representations from affective acoustic signals. Furthermore, the proposed WSM-MSN module, which integrates WST features with a CNN-HRNet multiscale fusion architecture, achieves additional performance gains. It exhibits robust and consistent performance across EMODB, RAVDESS, and eNTERFACE'05, significantly outperforming both conventional MFCC-based and baseline ScatNet methods in WAR and UAR. This underscores the superior capability of the WST-based front-end combined with the multiscale fusion back-end in processing affective acoustic signals.

Notably, on the more complex and conversational IEMOCAP dataset, while the WAR metric shows a marginal deficit, the module achieves a higher UAR. This phenomenon is primarily attributed to the imbalanced distribution of emotion categories within the IEMOCAP dataset, as well as the model's orientation toward the balanced optimization of recognition capability across

Table 2
Performance characteristics of MFCC, ScatNet, and WSM-MSN across different datasets.

Dataset	MFCC		ScatNet		WSM-MSN	
	WAR	UAR	WAR	UAR	WAR	UAR
EMODB (%)	58.93	54.03	74.40	71.30	**75.44**	**72.92**
RAVDESS (%)	36.74	34.77	50.00	48.50	**54.38**	**53.78**
IEMOCAP (%)	55.54	47.19	**60.41**	50.40	55.99	**54.84**
eNTERFACE'05 (%)	42.31	42.20	52.66	52.73	**56.12**	**56.10**

different categories during training. To address the above issues, optimization can be pursued from the following aspects: first, introducing class-balancing strategies into the loss function to balance the optimization intensity between majority and minority classes; second, fusing WST features with traditional acoustic features such as MFCCs to enhance the model's discriminative capability on majority classes; third, incorporating multimodal data information to improve comprehensive recognition performance through cross-modal information complementarity.

To further evaluate the model's discriminative ability across emotion categories, we analyze its confusion matrix. The categories considered include anger (Ang), boredom (Bor), disgust (Dis), fear (Fea), happiness (Hap), neutral (Neu), sadness (Sad), and surprise (Sur).

Figure 4 presents the emotion classification results across four benchmark datasets in confusion matrix form. Figure 4(a) indicates that the algorithm achieves the highest accuracy for anger, with overall higher performance on negative than positive emotions. Figure 4(b) shows the highest recognition rate for anger (70.31%), alongside robust performance for calmness and disgust, although some interclass confusion persists. Figure 4(c) reveals that while the recognition of the remaining three emotion categories is relatively balanced, the agitation category is frequently confused with anger and neutral. Figure 4(d) shows high accuracy for both anger/sadness and happiness/surprise pairs, highlighting the model's effectiveness in processing under-represented emotional categories. Collectively, these confusion analyses substantiate that the developed framework delivers precise and generalizable affective state classification, a cornerstone for deploying reliable audio-sensing modules in intelligent interactive systems.

4.5 Comparative experiments with existing algorithms

Table 3 shows the results for the WSM-MSN algorithm on the EMODB and RAVDESS datasets with those for existing methods. On the EMODB dataset, the proposed WSM-MSN method achieves a competitive performance of 75.44% WAR and 72.92% UAR, ranking among the top compared methods and second only to the state-of-the-art CQT-MSF with the DNN-SVM approach. On the RAVDESS dataset, the proposed algorithm attains 54.38% WAR and 53.78% UAR. The proposed method outperforms all compared methods in both WAR and UAR, demonstrating the highest overall recognition performance.

Table 4 shows the results on the IEMOCAP and eNTERFACE05 datasets. On the IEMOCAP dataset, the proposed method achieves 55.99% WAR and 54.84% UAR. While its WAR is slightly below that of the continuous wavelet transform (CWT)-based Conv2D-LSTM approach (57.31%), it attains a notably higher UAR, outperforming all compared methods in this metric.

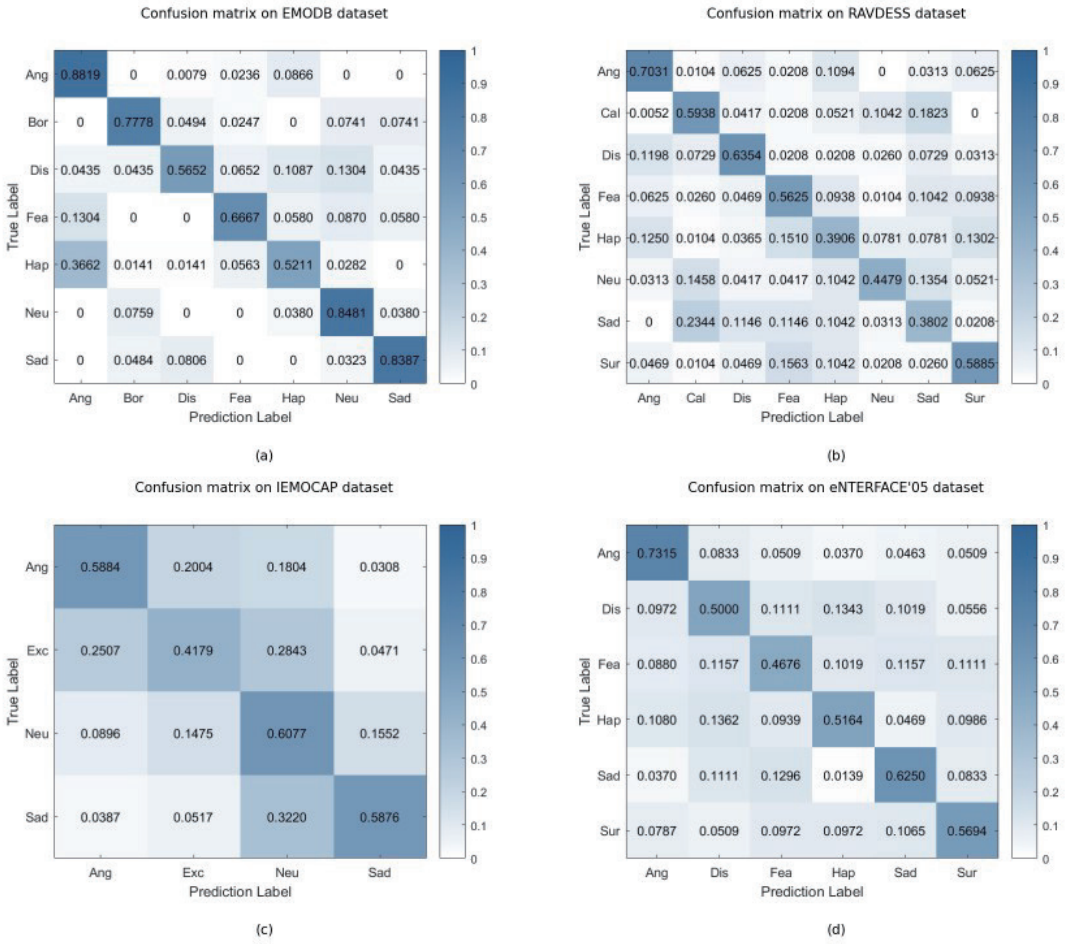

Fig. 4. (Color online) Confusion matrices of WSM-MSN.

Table 3
Results for methods related to WSM-MSN algorithm on EMODB and RAVDESS datasets.

Reference	Feature	Architecture	EMODB (%)		RAVDESS (%)	
			WAR	UAR	WAR	UAR
Avila *et al.*[23]	Modulation Spectrum	DNN	55.40	46.83	37.77	29.10
Liu *et al.*[24]	Interspeech 2009	Bi-LSTM	53.45	47.49	24.58	21.79
Singh *et al.*[8]	CQT	Conv2D-LSTM	65.69	60.45	46.94	44.15
Gerczuk *et al.*[10]	Mel Spectrum	ResNet	76.34	71.07	52.56	49.46
Singh *et al.*[25]	CQT-MSF	DNN-SVM	**79.86**	**76.17**	52.24	48.83
Our method	WSM	WSM-MSN	75.44	72.92	**54.39**	**53.78**

On the eNTERFACE’05 dataset, the proposed method obtains 56.12% WAR and 56.10% UAR. Although its WAR is lower than that of the CWT-Conv2D-LSTM model (59.05%), it achieves the highest UAR among all compared methods, demonstrating balanced and competitive overall performance.

The consistent performance gains across diverse acoustic datasets underscore the robustness and generalization capacity of the WSM-MSN framework, affirming its core attributes of noise robustness and cross-scenario adaptability.

Table 4
Results for methods related to WSM-MSN algorithm on IEMOCAP and eNTERFACE'05 datasets.

Reference	Feature	Architecture	IEMOCAP (%)		eNTERFACE'05 (%)	
			WAR	UAR	WAR	UAR
Singh *et al.*[8]	CQT	Conv2D	53.08	45.04	46.77	41.31
Singh *et al.*[8]	MFSC	LSTM + Attention	45.68	40.41	40.02	39.16
Singh *et al.*[8]	CWT	Conv2D-LSTM	**57.31**	49.94	**59.05**	54.80
Our method	WSM	WSM-MSN	55.99	**54.84**	56.12	**56.10**

4.6 Ablation experiment

To further investigate the contribution of each component, we conduct ablation experiments, with the results being presented in Table 5.

The ablation experiment results fully validate the necessity of each component in the proposed architecture. WSM-CNN achieves stable performance across all datasets, demonstrating that the WST front-end can effectively extract robust acoustic features. In contrast, WSM-HRNet exhibits a sharp performance drop, proving that the HRNet structure alone struggles to effectively learn emotion-discriminative information from raw signals. The complete WSM-MSN model, which integrates WST features with enhanced multiscale fusion architecture, achieves optimal performance across all four datasets. This indicates that the WST front-end provides high-quality input representations for the model, while the multiscale fusion architecture further exploits complementary information in the time–frequency domain. The synergy between these two components constitutes the superior performance of WSM-MSN. In summary, the ablation experiment confirms the complementarity and effectiveness of WST preprocessing and the multiscale fusion structure in emotion recognition tasks.

4.7 Model complexity analysis

To evaluate the practical deployment potential of the proposed model, we quantitatively analyze its parameter count, floating-point operations (FLOPs), and CPU inference latency, and compare them with those of typical models. The results are presented in Table 6.

Experimental results show that the proposed model has a total of only 0.62M parameters, accounting for 5.3% of those of ResNet18 and 82.3% less than for the lightweight MobileNetV2. In terms of computational complexity, the proposed model requires only 0.008G FLOPs per forward pass, which is merely 0.44% of that of ResNet18 and 97.5% lower than that of MobileNetV2. The model achieves an average CPU inference latency of 23.06 ms, which meets the real-time requirements of speech emotion recognition applications.

In summary, the proposed model demonstrates excellent performance in terms of parameter efficiency, computational complexity, and inference speed. With its exceptionally low computational cost and compact model size, it fully validates its practical deployment potential on resource-constrained devices.

Table 5
Results of the ablation experiment.

Dataset	WSM-CNN		WSM-HRNet		WSM-MSN	
	WAR	UAR	WAR	UAR	WAR	UAR
EMODB (%)	69.27	67.40	30.78	22.36	**75.44**	**72.92**
RAVDESS (%)	51.67	50.59	22.43	21.29	**54.38**	**53.78**
IEMOCAP (%)	55.94	54.46	39.61	34.88	55.99	**54.84**
eNTERFACE'05 (%)	53.64	53.60	23.54	23.49	**56.12**	**56.10**

Table 6
Results of model complexity.

Model	Params (M)	FLOPs (G)	CPU Latency (ms)
ResNet18	11.69	1.820	—
MobileNetV2	3.50	0.320	—
Our Method	0.62	0.008	23.06

5. Conclusions

We presented WSM-MSN, a module designed to enhance the robustness of emotion recognition from acoustic sensor signals. Confronted with the inherent challenge of environmental noise in microphone-captured data, the proposed approach synergistically integrates a WST-based preprocessing front-end with a multiscale neural architecture. The WST front-end standardizes raw sensor signals into stable, noise-invariant feature maps, while the subsequent network effectively captures and fuses both fine-grained local features and broad contextual dependences from these maps. Extensive evaluations on four benchmark datasets of affective acoustic signals substantiate the module's efficacy. The results demonstrate consistent performance improvements over conventional and contemporary benchmarks, confirming its superior generalization capability across diverse recording environments and emotional categories, and underscoring the module's potential for safety-aware human–robot interaction. Future research will evolve along two primary trajectories. First, we will pursue multimodal sensor fusion, integrating this audio module with visual processing streams from optical sensors to construct a unified cross-modal perception system, thereby boosting robustness in complex scenarios. Second, we will focus on algorithm-hardware co-design, optimizing the computational efficiency of the WST and network.

Acknowledgments

This work was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C03033) and National College Students' Innovative Entrepreneurial Training Program (No. 202410336048).

References

1 A. H. Al-Timemy, Y. Serrestou, R. N. Khushaba, S. Yacoub, and K. Raoof: IEEE Access **10** (2022) 107526. https://doi.org/10.1109/ACCESS.2022.3212146
2 F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss: Proc. Interspeech (2005) 1517. https://doi.org/10.21437/Interspeech.2005-446

3 S. R. Livingstone and F. A. Russo: PLoS One **13** (2018) e0196391. https://doi.org/10.1371/journal.pone.0196391
4 C. Busso, M. Bulut, and C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S. S. Narayanan: Lang. Resour. Eval. **42** (2008) 335. https://doi.org/10.1007/s10579-008-9076-6
5 O. Martin, I. Kotsia, B. Macq, and I. Pitas: Proc. 22nd Int. Conf. Data Engineering Workshops (ICDEW'06) (Atlanta, GA, USA, 2006) 8. https://doi.org/10.1109/ICDEW.2006.145
6 J. Wang, N. Ying, C. Zhu, Z. Liu, and Z. Cai: Telecommunications Science **35** (2019) 100.
7 M.R. Falahzadeh, E.Z. Farsa, A. Harimi, A. Ahmadi, and A. Abraham: IEEE Access **10** (2022) 112460.https://doi.org/10.1109/ACCESS.2022.3217226
8 P. Singh, S. Waldekar, M. Sahidullah, and G. Saha: Digital Signal Process. **130** (2022) 103712. https://doi.org/10.1016/j.dsp.2022.103712
9 M.R. Falahzadeh, F. Farokhi, A. Harimi, and R. Sabbaghi-Nadooshan: Circuits Syst. Signal Process. **42** (2023) 449. https://doi.org/10.1007/s00034-022-02130-3
10 M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller: IEEE Trans. Affective Comput. **14** (2023) 1472. https://doi.org/10.1109/TAFFC.2021.3135152
11 M. Mohan, P. Dhanalakshmi, and R.S. Kumar: Procedia Comput. Sci. **218** (2023) 1857. https://doi.org/10.1016/j.procs.2023.01.163
12 T. M. L. Flower and T. Jaya: Biomed. Signal Process. Control **93** (2024) 106201. https://doi.org/10.1016/j.bspc.2024.106201
13 H. Li, J. Li, H. Liu, T. Liu, Q. Chen, and X. You: Sensors **24** (2024) 5506. https://doi.org/10.3390/s24175506
14 K. Dabbabi and A. Mars: J. Syst. Sci. Syst. Eng. **33** (2024) 576. https://doi.org/10.1007/s11518-024-5607-y
15 J. H. Chowdhury, S. Ramanna, and K. Kotecha: Sci. Rep. **15** (2025) 11824. https://doi.org/10.1038/s41598-025-95734-z
16 P. Singh, G. Saha, and M. Sahidullah: Proc. 2021 29th European Signal Processing Conf. (EUSIPCO) (Dublin, Ireland, 2021) 131–135. https://doi.org/10.23919/EUSIPCO54536.2021.9615958
17 X. Y. Kek, C. S. Chin, and Y. Li: IEEE Access **10** (2022) 82185. https://doi.org/10.1109/ACCESS.2022.3196338
18 C. Sun, L. Ma, and H. Li: J. Signal Process. **39** (2023) 688. https://doi.org/10.16798/j.issn.1003-0530.2023.04.010
19 X. Yu and X. Li: Coal Science and Technology **52** (2024) 70.
20 W. Cheng, H. Zhang, and X. Gao: Information Countermeasure Technology **3** (2024) 51.
21 C. Sun, H. Li, and L. Ma: Front. Psychol. **13** (2023) 1075624. https://doi.org/10.3389/fpsyg.2022.1075624
22 X. Sun, Y. Gao, H. Lin, and H. Liu: Proc. ICASSP 2023 - 2023 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) (Rhodes Island, Greece, 2023) 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096309
23 A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk: IEEE Trans. Affective Comput. **12** (2021) 177. https://doi.org/10.1109/TAFFC.2018.2858255
24 Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao: Speech Commun. **139** (2022) 1. https://doi.org/10.1016/j.specom.2022.02.006
25 P. Singh, M. Sahidullah, and G. Saha: Speech Commun. **146** (2023) 53. https://doi.org/10.1016/j.specom.2022.11.005