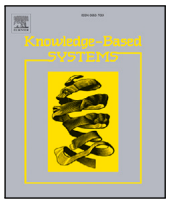
Knowledge-Based Systems 351 (2026) 116757

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Multi-feature wavelet attention network for audio deepfake detection ☆

Bo Wang [a], Rui Wang [b], Wei Wang [c,*], Zhongjie Ba [d]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, Liaoning, China
[b] Suyu Middle School, Suqian, 223800, Jiangsu, China
[c] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[d] School of Cyber Science and Technology, Zhejiang University, Hangzhou, 310027, Zhejiang, China

ARTICLE INFO

Keywords:
Wavelet transform
Automatic speaker verification
Audio deepfake detection
Self-supervised learning
Wav2Vec 2.0 XLSR

ABSTRACT

The abuse of synthetic speech poses significant threats to social security. However, existing detection models often struggle with generalization, especially when confronted with novel and unknown attacks. To address this challenge, we propose a novel method for audio deepfake detection called multi-feature wavelet attention audio deepfake detection (MWAADD). Specifically, we first extract self-supervised speech sequence features and emotional features using the wav2vec2 XLSR pre-trained model and an emotion recognition model. To better understand and capture these features, we introduce wavelet transform into deep neural networks, which enhances the model's ability to capture detailed information. Additionally, we combine this with an attention mechanism to focus on the most discriminative speech features. We further incorporate a Long Short-Term Memory (LSTM)-attention module to capture temporal dependencies and selectively focus on key features, ensuring the model highlights the most important clues. Experimental results show that our method outperforms existing state-of-the-art techniques, particularly in terms of generalization and robustness. Notably, on the ASVspoof 2019LA, ASVspoof 2021DF, and In-the-Wild datasets, our method achieves equal error rates of 0.18%, 2.55%, and 9.08%, respectively, demonstrating its superior detection performance.

1. Introduction

The advancements in speech synthesis technology have led to increasingly realistic synthesized speech, thereby presenting significant challenges in distinguishing it from genuine speech [1–6]. Criminals utilize the generated spoofed audio to commit crimes such as telephone fraud and extortion, which directly endangers the property safety of citizens [7–9]. At the same time, maliciously synthesized speeches are widely spread on social media, potentially undermining personal reputations and impacting social stability [10–13]. Furthermore, these forged speeches pose a serious threat to automated speaker verification (ASV) systems [14–16]. Therefore, the development of efficient audio deepfake detection technology has become a critical priority.

To reduce the risks associated with fake audio, researchers have proposed various detection approaches based on acoustic features and deep learning models. Conventional handcrafted features, such as linear frequency cepstrum coefficient (LFCC) [17], constant Q transform (CQT) [18], and Mel-frequency cepstrum coefficien (MFCC) [19], extract low-level time-frequency representations through signal processing techniques. Meanwhile, convolutional neural networks (CNNs) have been widely used due to their ability to learn complex patterns from audio [20–22]. Although traditional cepstral features have been widely used in spoof detection, they often fail to capture localized artifacts introduced across different frequency bands by advanced forgery algorithms, limiting robustness in real-world scenarios [23]. In addition, as handcrafted representations, they inherently lack the ability to encode high-level semantic cues, making it difficult to model the subtle characteristics of synthetic speech. Meanwhile, traditional CNN-based models also suffer from poor generalization, particularly under unseen or low-quality conditions [24]. This is partly due to the merging operations frequently adopted in CNNs, such as pooling or downsampling, which may lead to the loss of temporal information that is crucial for audio analysis.

Given the unknown forgery types and noise in real-world scenarios, and the limited generalization of prior methods, enhancing the robustness and adaptability of detection systems remains a critical challenge. To address these limitations, recent studies have turned to features extracted from self-supervised pre-trained models, such as wav2vec2 XLSR [25], which demonstrate strong generalization ability across

☆ This work is supported by the National Natural Science Foundation of China (No. 62372452), and by Youth Innovation Promotion Association CAS.
* Corresponding author.
E-mail addresses: bowang@dlut.edu.cn (B. Wang), wangrui11291224@163.com (R. Wang), wwang@nlpr.ia.ac.cn (W. Wang), zhongjieba@zju.edu.cn (Z. Ba).

https://doi.org/10.1016/j.knosys.2026.116757
Received 23 May 2025; Received in revised form 24 June 2026; Accepted 28 July 2026
Available online 3 August 2026

diverse domains [26]. Another practice is to incorporate emotional features into audio deepfake detection to enhance the discriminative effect. Since existing speech synthesis technology has difficulty in generating high-level speech features, such as emotional features [27–29]. In parallel, multi-scale analysis has emerged as a promising direction, which has proven effective in capturing inconsistencies across varying temporal and frequency resolutions. In particular, wavelet transform offers a powerful tool for such multi-scale analysis, enabling the extraction of localized frequency components that are often indicative of manipulation. Prior studies have shown that wavelet-based features are capable of revealing subtle artifacts and enhance generalization to unseen attacks [30–32].

In this paper, we propose a multi-feature wavelet attention neural network for audio deepfake detection (MWAADD), which integrates multi-features and multi-scale feature processing to enhance robustness under unknown attacks and noisy conditions. Specifically, this work utilizes the wav2vec2 XLSR (XLSR) and a speech emotion recognition (SER) pre-trained model to extract multi-features, effectively capturing both the acoustic sequence and emotional information. These features are fused and fed into the proposed multi-scale wavelet attention neural network (MWANN). Unlike conventional approaches that apply wavelet transforms directly on raw signals, we use wavelets as an analytical tool to analyze mutational structures in deep representations in a more fine-grained manner, which are key clues in synthesized speech. This approach retains the semantic expression of pre-trained features while introducing the multi-scale modeling capability of wavelets. To overcome the lack of focus in traditional wavelet decomposition, we introduce a self-attention mechanism after the transform to emphasize discriminative components, especially those in high-frequency bands where artifacts often manifest [33]. This attention-enhanced decomposition allows the model to adaptively focus on spoof-relevant features. The resulting embeddings are passed to an LSTM-attention classifier to model temporal dependencies and perform final classification. This architecture demonstrates superior generalization and robustness across multiple benchmark datasets, especially under low-quality and unseen spoofing conditions. The main contributions of this work are summarized as follows:

- Developed a novel detection method that synergistically integrates both acoustic sequences and emotional features, significantly enhancing the ability to identify spoofed speech.
- Designed an advanced classification framework that combines multi-scale attention mechanisms with the wavelet transformer, markedly improving the model's generalization and robustness.
- Sufficient experiments have been conducted on various public datasets and the results confirm that our proposed method exhibits outstanding detection capabilities.

This paper is organized as follows. Section 2 reviews the related works, Section 3 outlines the proposed detection method, Section 4 details the experimental setup, and Section 5 analyzes the experimental results in detail. Conclusions in Section 6.

2. Related work

With the rapid advancement of speech synthesis technologies, audio deepfake detection has gradually evolved from traditional handcrafted feature-based approaches to deep representation learning frameworks [34–37]. Representative studies have explored different strategies for improving detection performance. For instance, RawNet2 [38] directly learns discriminative representations from raw speech waveforms without relying on handcrafted frontends, whereas AASIST [39] employs integrated graph attention mechanisms to capture spectro-temporal dependencies for robust spoofing detection. Other studies have further incorporated pitch-related subband representations to enhance forgery-sensitive cues [40], systematically investigated the effectiveness of self-supervised pre-trained frontends [41], and improved robustness against noisy synthetic speech through knowledge distillation [3]. These advances demonstrate that robust feature representation learning has become a central research focus in modern audio deepfake detection. Building upon this trend, the following subsections review recent progress in self-supervised feature extraction and wavelet-based representation learning, which are most closely related to the proposed framework.

2.1. Self-supervised pre-trained feature extraction

Recent developments in large-scale pre-trained speech models have significantly promoted the application of self-supervised representations in audio deepfake detection. Trained on massive amounts of unlabeled speech data, these models are capable of learning rich contextual and phonetic representations with strong generalization ability. Recent benchmark studies have also systematically investigated different self-supervised frontends for speech spoofing countermeasures and demonstrated that pre-trained representations consistently outperform conventional handcrafted acoustic features under both in-domain and cross-domain evaluation settings.

Xiao et al. [42] proposed a dual-column Mamba architecture combined with the pre-trained wav2vec 2.0 model for spoofing detection, achieving significant improvements on challenging datasets. Similarly, Rosello et al. [43] employed Conformer networks with wav2vec 2.0 features, demonstrating strong performance in synthetic speech detection. Zhang et al. [44] fused XLSR and HuBERT representations to obtain more robust speech embeddings. Yang et al. [45] compared fourteen different speech representations and showed that features extracted from pre-trained models exhibit substantially better cross-dataset generalization than conventional handcrafted features. Wu et al. [46] further introduced self-supervised learning into adversarial defense for automatic speaker verification systems, demonstrating the robustness of pre-trained speech representations against adversarial attacks.

Overall, self-supervised pre-trained models have significantly improved feature representation and model generalization in audio deepfake detection. Nevertheless, these representations mainly focus on linguistic and acoustic information while overlooking certain high-level semantic cues, such as emotional characteristics [47,48]. Motivated by this observation, our method combines emotional features with contextual embeddings extracted from a self-supervised speech model, enabling the network to simultaneously capture high-level semantic information and deep phonetic representations. Such complementary feature fusion enhances discriminative capability while improving robustness against unseen synthetic attacks.

2.2. Wavelet transform for audio deepfake detection

Wavelet transform has attracted considerable attention owing to its excellent capability for multi-scale time-frequency analysis [49–51]. Leveraging this property, a number of studies have incorporated wavelet analysis into audio deepfake detection to enhance the characterization of subtle synthesis artifacts. Novoselov et al. [30] first introduced wavelet transform into audio deepfake detection,comparing the effectiveness of phase, MFCC, and wavelet-based semantic representations. Fathan et al. [23] further investigated multiple wavelet decomposition strategies on Mel-spectrograms and proposed WaveletCNN to improve spectral feature modeling. Gupta et al. [31] employed the continuous wavelet transform with generalized Morse wavelets to capture transient cues like signal bursts, which are key for detecting spoofed speech. Their study showed that Morse wavelet-based features outperformed short-time Fourier transform (STFT) and CQT in detection accuracy. Similarly, Avaiya et al. [32] introduced scattering transform to extract stable, multi-scale time-frequency representations, enhancing sensitivity to subtle signal inconsistencies and improving cross-domain robustness.

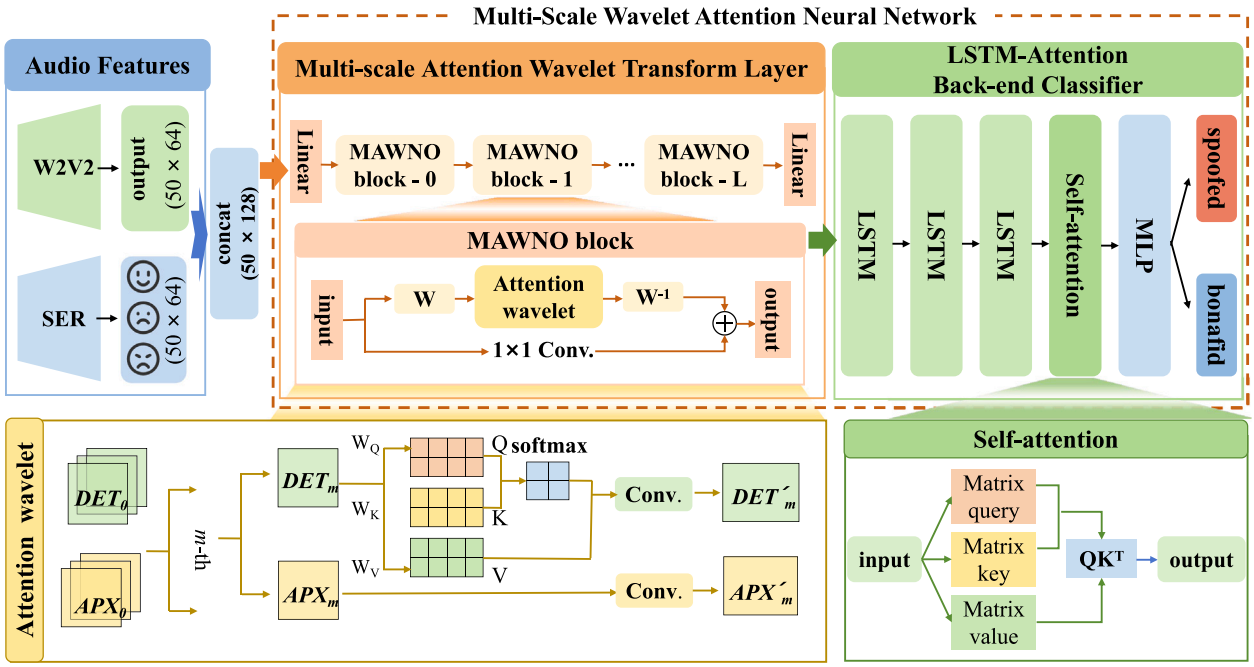


Fig. 1. The overall framework of MWAADD. This framework extracts semantic features from XLSR and SER. Then, it uses multi-scale attention wavelet transform layers to capture spoofing clues and performs final classification through LSTM-attention module. The light yellow box at the bottom represents the attention wavelet block and the green box represents the self-attention module. Here L is 4.

Although existing wavelet-based approaches have achieved promising performance, most of them apply wavelet transform directly to raw waveforms or spectrograms, which primarily capture low-level acoustic details while insufficiently exploiting high-level semantic representations. With the emergence of large-scale self-supervised speech models, pre-trained embeddings have become substantially richer and more robust than conventional handcrafted features. Therefore, rather than applying wavelet transform directly to the input speech signal, we introduce a wavelet-based attention mechanism into the embedding space of pre-trained representations to enhance temporal-frequency discrimination while preserving semantic information. Similar ideas have been successfully explored in computer vision, where WaveMix [52] and WaveMixSR [53] employ wavelet transforms as feature-space token mixers instead of operating on raw inputs. Inspired by these studies, our design combines the semantic richness of pre-trained representations with the multi-scale modeling capability of wavelet transform, leading to more discriminative and generalizable representations for audio deepfake detection.

3. Proposed method

In this session, we elaborate on the overall framework of the MWAADD, as shown in Fig. 1. The detection framework comprises feature extraction and the MWANN. Specifically, MWANN consists of multi-scale attention wavelet transform layers and LSTM-attention back-end classifiers. Together, these components provide a comprehensive approach for audio deepfake detection.

3.1. Overview

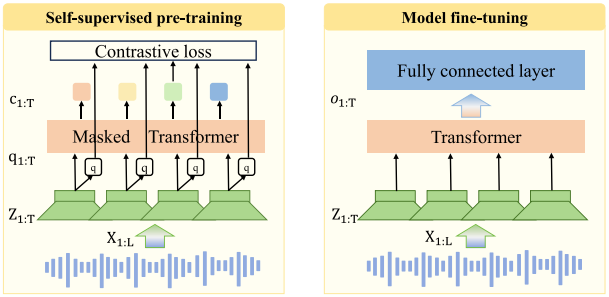
As demonstrated in Fig. 1, the SER module and the XLSR module receive the audio data simultaneously for fine-tuning, yielding emotional and sequence features. Subsequently, these extracted features are input into the multi-scale attention wavelet transform layers to capture more detailed embedding information. Then, MWAADD employs a LSTM-attention mechanism to concentrate on relevant characteristics across various time steps, thereby improving the overall feature representation capability. The output embeddings are subsequently processed by the multi-layer perceptron (MLP) for final classification. Overall, MWAADD integrates sequence and emotional features, employing wavelet attention, LSTM, and self-attention mechanisms, ultimately executing classification via fully connected layers.

3.2. Feature extractors

3.2.1. wav2vec2 XLSR

Based on wav2vec 2.0, XLSR is a large-scale model for learning cross-lingual speech representations. It has demonstrated strong performance in spoofed audio detection tasks. Its ability to capture rich phonetic and structural information makes it particularly suitable for identifying subtle artifacts and inconsistencies in synthesized speech. Fig. 2 illustrates the architecture of the network, which is composed of multi-layer CNNs and a transformer network. In the pre-training phase, the model undergoes training utilizing a contrastive loss function. XLSR transforms the input audio $\boldsymbol{X}_{1:L}$ into latent speech representations $\boldsymbol{Z}_{1:T}$ using a convolutional feature encoder network. The convolutional feature encoder network is composed of multiple CNN modules. The modules collaboratively process the input audio to extract its latent feature representation incrementally. At the same time, $\boldsymbol{Z}_{1:T}$ is discretized into a finite set of speech representations, denoted as $\boldsymbol{q}_{1:T}$. The feature $\boldsymbol{Z}_{1:T}$ is simultaneously masked and input into a transformer network to generate contextualized representations $\boldsymbol{c}_{1:T}$. In this study, we select the pre-trained Wav2Vec 2.0 XLSR 300M model. It was pre-trained with up to 2 billion parameters on 436K hours of publicly available speech audio across 128 languages.

Given that the pre-training phase utilizes only unlabeled bonafide audio, it is essential to fine-tune the detection system with both spoof and bonafide training data. As illustrated in Fig. 2(b), in the fine-tuning process, the original speech $\boldsymbol{X}_{1:L}$ is also input into the convolutional feature encoder network. The resulting feature vector $\boldsymbol{Z}_{1:T}$ is then passed into a transformer network, which produces the output sequence $\boldsymbol{o}_{1:T}$. We treat the output of the final transformer layer as the embedding feature, as it captures rich contextual and semantic information suitable for downstream classification tasks. The entire classification model is fine-tuned using the ASVspoof 2019 LA training data.



(a) Pre-training phase (b) Fine-tuning phase

Fig. 2. An overview of the wav2vec2 XLSR model. (a) Pre-training phase; (b) Fine-tuning phase. The key difference is that fine-tuning directly receives features from the Transformer without masking.

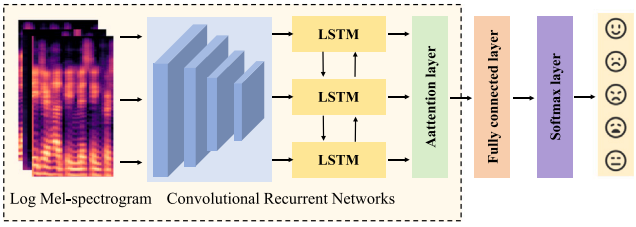


Fig. 3. The overall framework of SER. It extracts the log-mel spectrogram from the audio, which is then input into the convolutional recurrent neural networks. Finally, sentiment classification is performed using the fully connected layer. The light yellow box shows the emotional feature extraction part we used.

3.2.2. SER

Recent speech synthesis technology encounters significant difficulties in the precise generation of natural emotional expressions. Consequently, the extraction of emotional features can facilitate the effective classification of speech authenticity. The emotional features utilized in this study are derived from the 3D convolutional recurrent neural network proposed in the literature [54]. The network structure is illustrated in Fig. 3. To address the speech emotion recognition task, the author defines it as a multi-category classification problem, assuming the existence of N potential emotion categories in the speech. For the given speech signal $\boldsymbol{X}$, the output of the network is the emotion category set $E_{\boldsymbol{X}} \in \{e_1, e_2, \ldots, e_i, \ldots, e_n\}$, where e_i represents the specific emotion category. They proposed that input speech signal $\boldsymbol{X}$ needs to be preprocessed before entering the emotion classification network. The spectrum of $\boldsymbol{X}$ is initially computed using the STFT, followed by a logarithmic transformation of the STFT amplitude to produce a logarithmic spectrogram. Subsequently, the logarithmic spectrogram is processed through a series of 3D convolutional layers, followed by linear layers. After that, bidirectional long short-term memory networks (Bi-LSTM) are employed, along with an attention layer. Finally, the probability distribution of each emotion category is output through a softmax layer. To prepare for deepfake detection, we extract the output from the final attention layer, which captures the utterance-level sentiment features. The model is pre-trained on the IEMOCAP dataset [55] and then fine-tuned to adjust to the audio spoofing detection task. As a result, the extracted emotion vector could be used to recognize synthetic speech as well as categorize emotions.

3.3. Multi-scale attention wavelet neural network

The XLSR and SER features are first concatenated along the channel dimension, after which the fused representation is passed into the proposed multi-scale attention wavelet neural network. This network is composed of two main components: the multi-scale attention wavelet transform layer and the LSTM-attention back-end classifier.

3.3.1. Multi-scale attention wavelet transform layer

The wavelet transform is an effective multi-scale analytic tool that has typically been used to divide raw signals into components with varying temporal and frequency resolution. Instead of applying wavelets to raw audio, we use this capability within the neural embedding space, specifically the feature representations generated by pre-trained models. By applying the wavelet transform to these embeddings, we enable multi-resolution decomposition of high-level semantic features, which facilitates the detection of subtle, localized artifacts that may be blurred or lost in traditional feature spaces. This design leverages the semantic richness of neural embeddings while benefiting from the multi-scale analysis capability of wavelets. To further enhance the model's ability to identify spoof-specific cues, we propose the multi-scale attention wavelet neural operator (MAWNO), which adopts a structured pipeline composed of discrete wavelet transform (DWT), the self-attention module, and inverse discrete wavelet transform (IDWT). First, the DWT enables multi-scale decomposition along the temporal axis of neural embeddings, which allows the model to capture fine-grained temporal variations that frequently characterize synthetic speech. Second, conventional wavelet filtering is static and lacks adaptability [56]. To overcome this, we introduce a self-attention mechanism to selectively refine the high-frequency wavelet coefficients, enabling the model to dynamically focus on regions that are more likely to contain discriminative artifacts. Finally, the IDWT reconstructs the modified frequency components back into the temporal feature space, effectively preserving local discriminative. Through this integration of multi-scale analysis and deep semantic modeling, the MAWNO module significantly boosts the model's discriminative capacity to detect artifacts in challenging audio deepfake scenarios.

Fig. 1 illustrates the overall structure of the multi-scale attention wavelet transform layer. First, the input features are processed through the linear layer to map the features into a higher-dimensional space. This process enables the model to capture complex patterns and relationships within the audio. The transformed features are subsequently processed by L stacked MAWNO blocks, and then restored to their original dimensionality using another linear layer. Each MAWNO block consists of three core steps: wavelet decomposition, attention-enhanced processing, and inverse wavelet reconstruction. As shown in Fig. 1, the input $I(\boldsymbol{X})$ is first decomposed into multi-scale coefficients using the DWT, denoted by $\mathcal{W}$. Specifically, DWT decomposes the input feature sequence into m levels (APX_0, DET_0, …, APX_m, DET_m). Contrary to classical assumptions, the lower-level coefficients (e.g., DET_0) capture high-frequency details that may contain critical manipulation traces as well as noise, while higher-level coefficients correspond to progressively lower-frequency bands. In our wavelet attention Module, we do not impose a priori assumptions on which sub-bands are informative. Instead, we apply a self-attention mechanism to the high-frequency coefficients at the final decomposition level (e.g., DET_m) to adaptively emphasize regions likely to contain subtle manipulation artifacts. Subsequently, a convolutional operation further enhances discriminative patterns before recombining all sub-bands via the IDWT, yielding a representation that integrates complementary information across scales. The self-attention mechanism is applied to the DET_m using the following formula:

$$\begin{aligned} \boldsymbol{Q} &= \mathrm{DET}_m \cdot \boldsymbol{W}_Q, \\ \boldsymbol{K} &= \mathrm{DET}_m \cdot \boldsymbol{W}_K, \\ \boldsymbol{V} &= \mathrm{DET}_m \cdot \boldsymbol{W}_V, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are learnable parameter matrices that define the attention formula.

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\top}}{\sqrt{d}}\right) \cdot \boldsymbol{V}. \tag{2}$$

By combining Eqs. (1) and (2), DET'_m is simply obtained:

$$\mathrm{DET}'_m = \mathrm{Conv1D}\left(\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\right). \tag{3}$$

Meanwhile, to enhance the spoof-awareness of the low-frequency components extracted via wavelet decomposition, the approximation coefficients APX_m are processed through a convolutional transformation to yield APX'_m. These modified coefficients are then recombined using the IDWT to reconstruct a refined features that retain both the original resolution and enhanced spoof-awareness. In parallel, the input $I(\boldsymbol{X})$ is processed through a 1×1 convolution path. This branch facilitates cross-channel feature fusion and introduces nonlinearity, which serve as a supplementary path to avoid the loss of useful original features. The final output $\boldsymbol{O}(\boldsymbol{X})$ is obtained by concatenating both branches. This dual-branch structure promotes a richer and more robust feature representation by effectively integrating the enhanced features with the original input features, thereby improving the model's generalization and spoof detection performance.

In our implementation, we adopt the Daubechies 6 (db6) wavelet family for both DWT and IDWT operations. This family offers a good trade-off between time localization and frequency smoothness, which is crucial for capturing transient anomalies introduced by spoofing techniques. Overall, the MAWNO module equips the model with the ability to analyze speech signals across multiple resolutions while adaptively focusing on discriminative patterns, thereby improving robustness against unseen and noisy forgeries.

3.3.2. LSTM-attention back-end classifier

The resulting feature vectors are then processed by LSTM models. LSTM can extract higher-level features from speech data [57]. To improve the model's processing of temporal information, we add an attention mechanism to improve performance by understanding context and dependencies. Specifically, [58,59] demonstrate the efficacy of self-attention in language modeling and enhancement. These findings highlight the powerful potential of attention mechanisms in speech signal processing. Let the output of the multi-scale attention wavelet neural layers be denoted as $\boldsymbol{O} \in \mathbb{R}^{B\times T\times F}$, where B is the batch size, T is the frame number, and F is the feature dimension. The equations for the LSTM-Attention module are given as follows:

$$\boldsymbol{O} \in \mathbb{R}^{B\times T\times F}, \tag{4}$$

$$\boldsymbol{h}_t = \mathrm{LSTM}(\boldsymbol{O}[t-1], \boldsymbol{O}[t]), \tag{5}$$

$$\boldsymbol{O}_{\mathrm{lstm+att}} = \sum_{t=1}^{T} \alpha_t \boldsymbol{h}_t, \tag{6}$$

where the attention weight α_t is calculated by:

$$\alpha_t = \frac{\exp\left(\boldsymbol{w}_a^{\top}\boldsymbol{h}_t\right)}{\sum_{k=1}^{T}\exp\left(\boldsymbol{w}_a^{\top}\boldsymbol{h}_k\right)}, \tag{7}$$

where $\boldsymbol{w}_a$ is a learnable parameter vector.

Eq. (5) represents the recursive formulation of the LSTM module, where the input feature at the tth time step, denoted as $\boldsymbol{O}[t]$, is combined with the previous state $\boldsymbol{O}[t-1]$ to compute the new hidden state $\boldsymbol{h}_t$. This formula describes the core mechanism of LSTM, which processes sequence data by leveraging the temporal dependencies between time steps. Eq. (6) denotes the final output resulting from the combination of LSTM and the attention mechanism. Specifically, $\boldsymbol{O}_{\mathrm{lstm+att}} \in \mathbb{R}^{B\times T\times F}$ denotes the aggregated sequence output, obtained by computing a weighted summation over all LSTM hidden states $\boldsymbol{h}_t$. The attention weights α_t are dynamically computed to reflect the relative importance of each time step in contributing to the final representation. To obtain optimal classification performance, the attention-enhanced features are passed through a three-layer multilayer perceptron with batch normalization and dropout applied to reduce overfitting. The fully connected process is described as:

$$\boldsymbol{x}_1 = \boldsymbol{W}_1 \cdot \boldsymbol{O}_{\mathrm{lstm+att}} + \boldsymbol{b}_1, \tag{8}$$

$$\boldsymbol{x}_2 = \boldsymbol{W}_2 \cdot \mathrm{Dropout}(\boldsymbol{x}_1) + \boldsymbol{b}_2, \tag{9}$$

$$\boldsymbol{y} = \boldsymbol{W}_3 \cdot \mathrm{Dropout}(\boldsymbol{x}_2) + \boldsymbol{b}_3. \tag{10}$$

Here, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ are learnable weight matrices, and $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$ are the corresponding bias vectors. The final output $\boldsymbol{y}$ represents the classification logits used for prediction.

4. Experiments

4.1. Experimental settings and metrics

4.1.1. Datasets

To assess the proposed method, we conducted experiments on multiple public audio spoofing datasets. These datasets generate forged audio using a variety of text-to-speech (TTS) and voice conversion (VC) algorithms, as well as their hybrids. This study utilized the ASVspoof 2019LA (19LA) training dataset [60] for training and assessed performance using the 19LA evaluation dataset, ASVspoof 2021LA (21LA) [61], ASVspoof 2021DF (21DF) [61], and an unseen dataset called In-the-Wild (ITW) [62]. Table 1 presents the detailed statistics of the experimental datasets, which are described as follows. The 19LA dataset, based on VCTK [63], includes speech forged by 19 TTS and VC algorithms. It integrates spoofing attacks created by 13 different algorithms, identified as A07 to A19. Among these, there are 11 unknown spoofing algorithms and 2 known algorithms (A16 and A19, using the same methods as A04 and A06). 21LA, derived from 19LA, simulates real-world transmission scenarios with varying codecs, bit rates, and sampling rates. 21DF is part of ASVspoof 2021, featuring spoofed audio from over 100 TTS and VC systems across three source domains (VCTK, VCC2018 [64], VCC2020 [65]), along with diverse compression schemes. The ITW dataset consists of real-world English audio collected from online media, including public speeches with various background conditions such as noise, reverberation, and streaming artifacts.

4.1.2. Evaluation metrics

This paper evaluates the effectiveness of the model and its impact on the ASV system using equal error rate (EER) and minimum tandem detection cost function (mt-DCF) as the primary indicators in the two datasets 19LA and 21LA. However, in the 21DF and ITW datasets, which focus exclusively on forged voice detection without considering the integration with the ASV system, the evaluation metric is limited to the EER.

EER. The EER is a widely used metric in audio deepfake detection. It refers to the error rate at the threshold where the false acceptance rate (FAR) equals the false rejection rate (FRR). A lower EER indicates better detection performance. Let θ denote the decision threshold, then:

$$\mathrm{EER} = P_{FAR}(\theta) = P_{FRR}(\theta). \tag{11}$$

mt-DCF. To assess the reliability of detection systems in practical ASV scenarios, the mt-DCF is introduced. It jointly considers detection errors and their cost impact on the ASV system. Defined as:

$$\text{mt-DCF} = \min\{\beta P_{FRR}(\theta) + P_{FAR}(\theta)\}, \tag{12}$$

where β reflects the relative cost and prior probabilities in a specific ASV configuration. A lower min t-DCF indicates better performance in real-world integration with speaker verification systems.

4.1.3. Implementation details

During the training phase, input audio was preprocessed at a sampling rate of 16 kHz, and the first 64,600 samples were selected through truncation or repetition. All models were trained using a batch size of 12 and optimized with the Adam optimizer, and weight decay of 10^{-4}. The loss function utilized was cross entropy loss, and the learning rate was set to 1×10^{-6}. Training was conducted for 100

Table 1
Summary of the dataset and detailed statistics.

Datasets	Genuine	Spoofed	Total
Train	2580	22,800	25,380
Dev	2548	22,296	24,844
Eval(19LA)	7355	63,882	71,237
Eval(21LA)	18,452	163,114	181,566
Eval(21DF)	22,617	589,212	611,829
Eval(ITW)	19,963	11,816	31,779

epochs with a warm-up period of 3 epochs, where only the training loss was monitored. Early stopping was implemented with a patience of 10 epochs. The entire framework was developed in Python 3.8 with PyTorch 1.8, and the wavelet transform operations were implemented using the pytorch_wavelets library. All experiments were carried out on a Linux operating system equipped with an NVIDIA GeForce RTX 4090 GPU and 128 GB RAM, with CUDA version 12.2 for GPU acceleration. To ensure temporal alignment between the two feature streams, we perform frame rate synchronization. The raw outputs from XLSR and SER differ in temporal resolution due to differences in architecture and preprocessing. To resolve this, both embeddings are passed through fixed max-pooling operations followed by BatchNorm1d over a unified frame number of 50. This normalizes both views to the same temporal resolution, allowing for concatenation along the feature dimension and enabling synchronized multi-feature learning. We used 4 stacked MAWNO Blocks to construct the multi-scale attention wavelet transform layer. During evaluation, t-SNE plots were generated to visualize feature embeddings, and normalized confusion matrices were plotted to assess classification performance. All the experiments are conducted on the same platform with a single NVIDIA GeForce RTX 4090 GPU.

5. Experimental results

5.1. Comparative experiments

5.1.1. Experimental results on ASVspoof 2019LA

First, we conducted tests on the evaluation set of 19LA. The results of the experiment are presented in Table 2. We assessed the detection models utilizing manual features, emotional features, original speech, and various pre-trained features as front-end features. Ref. [66] serves as the baseline model for ASVspoof 2019. The LFCC features of the speech signal are extracted and subsequently input into the GMM back-end classifier. The EER calculated for 19LA is 8.09%, while the mt-DCF is 0.2116. The effectiveness of this detection model is unsatisfactory. The performance of the two counterfeiting algorithms, A10 and A19, is notably inferior. Subsequently, Conti et al. [28] evaluated the detection model utilizing emotional features in the front-end and a random forest (RF) classifier in the back-end. Its detection performance on 19LA is better than that of [66], with an EER of 7.14%, which is 0.95% lower. Graph neural networks were introduced into speech detection by Jung et al. [39]. The EER on 19LA is 0.82%. The proposed lightweight graph neural network attained an EER of 0.99%. In comparison to traditional neural networks, graph neural networks demonstrate superior classification capabilities. Tak et al. [67] integrated features from XLSR pre-trained model with graph neural networks, demonstrating significantly enhanced classification performance. Their research incorporated data augmentation techniques. For a fair comparison, their model was retrained without augmentation, achieving an EER of 0.25% on the 19LA dataset.

Subsequently, we chose the detection network that performed the feature fusion. The pre-trained features XLSR and WavLM were fused, and the resulting features were input into LCNN for classification [44]. In comparison to graph neural networks, the detection performance of the simple convolutional network LCNN was diminished and struggle to adequately capture the features of XLSR and WavLM. In conclusion, pre-trained features, developed using extensive real-world data, demonstrate superior detection capabilities compared to handcrafted features. It is also equally essential to design a network that makes full use of these pre-trained features. We integrated the pre-trained features of XLSR with SER features and conducted back-end classification using a multi-scale attention wavelet neural network. The experimental results indicate that the proposed method demonstrates effective detection performance. The EER and mt-DCF for 19LA are 0.18% and 0.005, representing reductions of 0.62% and 0.016 compared to the previous state of the art. This demonstrates that our proposed MWANN can effectively and comprehensively understand the pre-trained features and apply them to the classification of genuine and spoofed audio.

5.1.2. Experimental results on ASVspoof 2021LA and ASVspoof 2021DF

In order to assess the detection performance of the proposed system in real-world scenarios, we conducted experiments on the 21LA and 21DF datasets. The experimental results of specific attack types were detailed. L03, L04, L05, L06, and L07 in 21LA correspond to the pstn, g722, ulaw, gsm, and opus codecs, respectively. D02, D03, D04, D05, and D06 in 21DF correspond to the low_mp3, high_mp3, low_m4a, high_m4a, and low_ogg compression formats, respectively. The compression scenario poses a considerable challenge for the audio spoofing detection task. Table 3 shows the detection performance under various attack modes in detail. Classical classification networks may struggle to show better detection performance. For instance, the TSSDNet [68] method demonstrates strong performance against certain forgery attacks (notably L04 and L05), but its overall detection efficacy across the complete 21LA and 21DF datasets is suboptimal. Rawnet2 [38], while effective as a classification network, demonstrates insufficient detection performance when exposed to speech subjected to varying transmission noise and compression coding conditions. Furthermore, while the graph neural network AASIST demonstrates strong performance on the 19LA dataset [39], it exhibits limitations in detecting speech that has been compressed or influenced by the transmission channel.

By fine-tuning the pre-trained XLSR and applying it to the forgery detection task, the detection performance is significantly improved. XLSR employs millions of hours of speech data during the pre-training phase to enhance the analysis and feature extraction of speech signals. [67] employs XLSR as the front-end feature extractor, while the back-end utilizes the advanced graph neural network AASIST. The training process enhances model performance via data augmentation. To ensure fairness, the data augmentation component was eliminated from the comparative experiment. As indicated in Table 3, the model proposed by [67] attained the EER of 8.13% and 7.36% for the 21DF and 21LA datasets [27], respectively. In comparison to prior state-of-the-art techniques, the EER shows a reduction of 12.12% and 5.46% for the 21DF and 21LA datasets, respectively.

Conversely, we integrated pre-trained XLSR features with emotional features and employed the MWANN. We investigated the application of wavelet transform within deep neural networks for forgery detection tasks, focusing on the enhancement of speech detail capture, which resulted in superior classification outcomes. In the 21DF and 21LA datasets, our model attained an EER of 2.55% and 4.15%, respectively, representing reductions of 5.58% and 3.21% compared to the previous best model, indicating superior detection performance.

5.1.3. Experimental results on In-the-Wild

Finally, We also conducted cross-database experiments on the ITW dataset. The attack types of the ITW dataset remains invisible to the model, and the dataset exhibits substantial variation in source domains, encompassing various recording devices, reverberation conditions, and background noise. Due to the diversity of these source domains, traditional detection networks exhibit inadequate efficiency on ITW. Table 4 indicates that the EER of the model on the ITW dataset is approximately 30% when employing original input speech. Detection algorithms dependent on raw audio exhibit limited generalization and robustness

Table 2
Comparison of detection performance based on EER(%) and mt-DCF on the ASVspoof 2019LA evaluation dataset.

Feature	Back-end	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	EER	mt-DCF
LFCC	GMM [66]	12.86	0.37	0.00	18.97	0.12	4.92	9.57	1.22	2.22	6.31	7.71	3.58	13.94	8.09	0.2116
Emotion	RF [28]	0.22	3.42	0.11	0.45	0.33	0.19	0.19	0.42	0.38	0.75	20.51	17.2	6.24	7.14	0.1691
Raw audio	AASIST [39]	0.47	0.41	0.00	0.79	0.16	0.67	0.12	0.16	0.51	0.65	1.28	2.65	0.67	0.82	0.0272
Raw audio	AASISTL [39]	0.61	0.20	0.00	1.06	0.16	0.77	0.20	0.06	0.61	0.65	1.91	3.03	0.69	0.99	0.0317
XLSR	AASIST [67]	0.06	0.06	0.02	0.40	**0.10**	0.14	0.00	0.06	0.24	**0.06**	0.37	0.84	0.35	0.25	0.0071
XLSR+WavLM	LCNN [44]	0.69	1.04	0.76	1.79	0.76	0.76	0.68	1.69	2.28	0.71	0.81	1.85	4.78	1.67	0.0537
XLSR+SER	MWANN	**0.02**	**0.06**	**0.00**	**0.32**	0.22	**0.02**	**0.00**	**0.02**	**0.02**	0.08	**0.24**	**0.19**	**0.01**	**0.18**	**0.0050**

Table 3
Robust performance comparison results based on EER(%) and mt-DCF on the ASVspoof 2021LA dataset, and comparison of detection performance based on EER(%) on the ASVspoof 2021DF dataset.

Feature	Model	21LA						21DF						
		L03	L04	L05	L06	L07	EER	mt-DCF	D02	D03	D04	D05	D06	EER
Raw audio	TSSDNet [68]	36.94	3.42	3.10	11.20	4.92	13.86	0.5743	36.31	35.77	35.34	35.36	24.72	29.98
Raw audio	Rawnet2[38]	16.72	6.41	6.33	10.66	7.95	9.50	0.4257	27.63	27.49	26.72	27.23	18.74	22.38
Raw audio	AASIST [39]	13.82	7.89	6.32	9.55	12.14	10.91	0.4931	24.75	23.88	23.73	24.20	18.49	20.96
Raw audio	AASISTL [39]	19.30	7.79	6.03	11.76	11.88	12.82	0.5468	24.12	23.79	23.46	23.47	17.40	20.25
XLSR	AASIST [67]	11.76	1.50	3.15	8.61	4.37	7.36	0.3769	8.99	9.74	8.63	8.99	7.50	8.13
XLSR+SER	MWANN	**6.54**	**0.87**	**1.44**	**5.00**	**2.93**	**4.15**	**0.2900**	**4.73**	**2.11**	**2.07**	**1.97**	**2.08**	**2.55**

Table 4
Generalization comparison results of detection performance based on EER (%) on the In-the-Wild dataset.

Features	Model	EER
Raw audio	TSSDNet [68]	34.16
Raw audio	AASIST [39]	43.50
Raw audio	AASISTL [39]	44.64
XLSR, WavLM, Hubert	Fusion (concat) [45]	24.27
XLSR	AASIST [67]	18.67
XLSR+SER	MWANN	**9.10**

in cross-database evaluations, struggling to identify previously unseen attack types.

The implementation of large pre-trained models has effectively improved this problem. However, traditional neural networks encounter challenges in processing pre-trained features, making it difficult to fully extract the effective information contained within these features. Just demonstrated as [45], they employed three widely used pre-trained models for feature extraction, but the lack of an appropriate classifier resulted in a final EER of 24.27% on the ITW dataset. Much improved results when using graph neural networks [67]. Consequently, designing a more robust back-end classification network is essential. Experimental results show that the EER of the proposed MWANN on the ITW dataset is only 9.10%, which is 9.57% lower than the previous optimal method, demonstrating the effectiveness of our method. It is demonstrated that MWAADD has significant advantages in handling cross-database data and addressing source domain diversity.

5.2. Ablation experiment

To assess the necessity of each module in the proposed WMAADD, we performed ablation experiments using the 19LA, 21LA, and 21DF datasets. Table 5 shows the detailed results of each model. We first evaluate the impact of different front-end features. Specifically, we feed the XLSR features into the proposed MWANN architecture, denoted as Model I. Experimental results demonstrate that fine-tuning the XLSR pre-trained model enables Model I achieves superior detection performance across all four datasets, attributed to the strong generalization capability of the XLSR features. Then, we replace the input with SER features alone, yielding Model II, and the results show that rely solely on emotional features resulted in a sharp decline in detection performance, highlighting the emotion model's limited ability to process speech features compared to the XLSR pre-trained model. Compared

Table 5
Ablation experiment results measured by EER (%) and mt-DCF on the 19LA, 21LA, and 21DF datasets. The compared models are defined as follows: Model I uses fine-tuned XLSR features with MWANN; Model II uses SER features with MWANN; Model III concatenates XLSR and SER features followed by a fully connected layer; Model IV applies the LSTM-attention module to the concatenated features; Model V integrates the concatenated features into the proposed multi-scale attention wavelet transform layer. The final MWAADD model concatenates XLSR and SER features and feeds them into the MWANN classifier. For clarity, all XLSR settings refer to fine-tuned weights, except where explicitly noted as frozen (see Table 6).

Model	2019LA		2021LA		2021DF
	EER	mt-DCF	EER	mt-DCF	EER
Model I	0.35	0.0118	5.21	0.3180	6.75
Model II	14.09	0.3636	17.85	0.5110	47.42
Model III	17.36	0.2900	14.67	0.4700	6.49
Model IV	0.24	0.0070	8.62	0.4300	3.73
Model V	0.37	0.0120	4.89	0.3200	3.43
MWAADD	**0.18**	**0.0050**	**4.15**	**0.2900**	**2.55**

with WMAADD and Model I, we discovered that, while emotional characteristics alone may not improve forgery detection, combining emotional features with XLSR substantially improves detection results by providing complementing high-level semantic cues that XLSR may not fully capture.

Subsequently, we conduct ablation studies on each component within MWANN to assess their individual contributions. First, we defined the baseline model without the multi-scale attention wavelet transform layer and the LSTM-attention module as Model III, and only input the fusion of XLSR and SER features into the fully connected layer. We deliberately chose a simple fully connected backend as a baseline to evaluate raw feature concatenation without additional modeling. The fair comparison using the same MWANN backend is presented in the final MWAADD model. The EER of Model III on the 19LA dataset is 17.36%. Interestingly, the detection effect of this model on the 21LA and 21DF datasets is better than that of 19LA, indicating that it has a certain generalization ability. This may be attributed to the robust feature extraction capabilities of the XLSR pre-trained model.

Based on Model III, we added the LSTM-attention module and built Model IV. By introducing this module, the performance of the model has been significantly improved. On the 19LA, 21LA, and 21DF datasets, the EER is reduced by 17.12%, 6.05%, and 2.76%, respectively. Additionally, in the 19LA and 21LA datasets, the mt-DCF is reduced by 0.283 and 0.04, respectively. The findings indicate that the

Table 6
Results of module comparison experiments in terms of EER (%) and mt-DCF on the 19LA, 21LA, and 21DF datasets. The compared models are defined as follows: Model I: XLSR (frozen) + SER features input to MWANN, Model II: Attention-based XLSR + SER features input to MWANN, Model III: XLSR + SER features with Haar wavelet replacing db6 in MWANN, Model IV: XLSR + SER features with db4 wavelet replacing db6 in MWANN, Model V: XLSR + SER features with LSTM module replaced by GRU in MWANN and Model VI: XLSR + SER features with LSTM module replaced by ResNet in MWANN.

Model	2019LA		2021LA		2021DF
	EER	mt-DCF	EER	mt-DCF	EER
Model I	8.95	0.2313	17.53	0.5535	16.97
Model II	0.19	0.0061	7.45	0.3949	5.31
Model III	0.54	0.0184	9.79	0.4413	5.31
Model IV	0.37	0.0096	7.01	0.3725	3.27
Model V	0.26	0.0088	7.05	0.3642	3.92
Model VI	0.24	0.0069	7.32	0.3912	6.25
MWAADD	**0.18**	**0.0050**	**4.15**	**0.2900**	**2.55**

LSTM-attention backend is capable of effectively processing pre-trained features, thereby improving the model's performance in speech forgery detection.

Model V was constructed following the incorporation of the multi-scale attention wavelet Transform layer into Model III. Experimental results indicate a notable enhancement in the detection performance of Model V, with an EER of 0.37%, 4.89%, and 3.43% on the 19LA, 21LA, and 21DF datasets, respectively. In comparison to Model III, the EER decreases by 16.99%, 10.52%, and 3.19%, respectively. The multi-scale attention wavelet transform layer effectively utilizes pre-trained features, thereby enhancing detection performance significantly. It is worth noting that Models IV and V were designed to illustrate the individual contributions of the LSTM-attention module and the multi-scale attention wavelet transform layer, rather than to serve as a direct pairwise comparison. The integration of both modules is evaluated in our final model.

Finally, we combined the MWANN with pre-trained features to construct the final MWAADD model. MWANN consists of the LSTM-attention module and the proposed multi-scale attention wavelet transform layer, which together provide complementary modeling capacity. As shown in the last row of Table 5, MWAADD shows excellent performance on all three datasets. The EER achieved on the 19LA, 21LA, and 21DF datasets are 0.18%, 4.15%, and 2.55%, respectively, while the mt-DCF for the 19LA and 21LA datasets are 0.005 and 0.29, respectively. The proposed model demonstrates strong detection capabilities on the high-quality 19LA dataset, as well as notable generalization and robustness on the more challenging datasets, 21LA and 21DF. Overall, our experiments verified the effectiveness of the each pre-trained features, multi-scale attention wavelet transform layer and LSTM-attention modules, which cooperate with each other to significantly improve the performance of the speech forgery detection system, especially under compressed or transmission conditions. It shows good robustness and wide adaptability.

5.3. Comparative analysis of various components in proposed framework

In order to verify the irreplaceability and superiority of the core modules in the proposed MWAADD network, we conducted comprehensive module comparison experiments. The detailed experimental results are shown in Table 6.

To assess the impact of fine-tuning the XLSR encoder, we construct Model I by freezing the XLSR extractor and fusing its outputs with SER into the MWANN framework. Experimental results indicate that the sharp decline in detection performance caused by freezing XLSR highlights the importance of fine-tuning in enabling the model to effectively adapt to the audio deepfake detection.

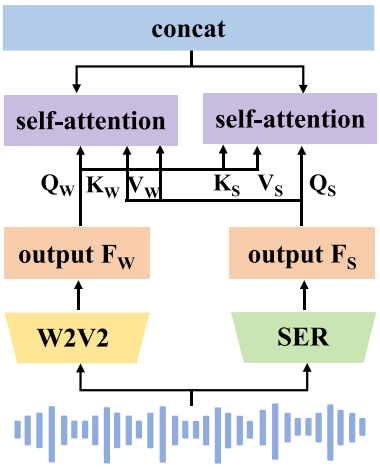


Fig. 4. Attention-based feature fusion framework. The original speech is input into wav2vec 2.0 (W2V2) and the speech emotion recognition (SER) module, which extract their respective features. These features are then fused using a self-attention mechanism to obtain the final feature representation.

This work employs a direct fusion method for processing front-end features while also investigating the alternative of feature fusion via the attention mechanism. Model II in Table 6 employs the attention mechanism (ATT-fusion) to integrate XLSR with SER features, subsequently feeding the combined features into the MWANN back-end network.

Fig. 4 illustrates the structure of attention-based fusion, where the query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) vectors are defined as follows. For wav2vec2 XLSR features, we set $\boldsymbol{Q}_w = \boldsymbol{F}_w$, $\boldsymbol{K}_w = \boldsymbol{F}_s$, and $\boldsymbol{V}_w = \boldsymbol{F}_s$. For emotion-based features, we set $\boldsymbol{Q}_s = \boldsymbol{F}_s$, $\boldsymbol{K}_s = \boldsymbol{F}_w$, and $\boldsymbol{V}_s = \boldsymbol{F}_w$, where $\boldsymbol{F}_w$ and $\boldsymbol{F}_s$ denote the features extracted from wav2vec2 XLSR and SER modules, respectively. The attention-based outputs are computed using the standard scaled dot-product attention mechanism:

$$\boldsymbol{O}_w = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_w \boldsymbol{K}_s^{\top}}{\sqrt{D}}\right)\boldsymbol{V}_s, \tag{13}$$

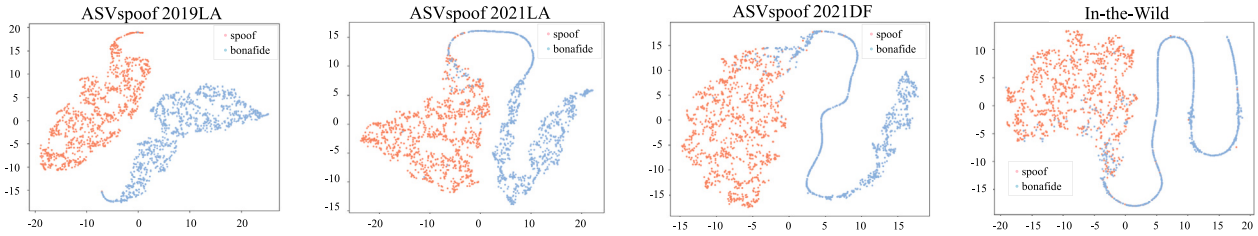
$$\boldsymbol{O}_s = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_s \boldsymbol{K}_w^{\top}}{\sqrt{D}}\right)\boldsymbol{V}_w, \tag{14}$$

where D is the dimensionality of the key vectors, set to 128. The final representations $\boldsymbol{O}_w$ and $\boldsymbol{O}_s$ are concatenated and subsequently passed to the backend classifier.

To investigate the influence of different wavelet bases on model performance, we replace the default wavelet (db6) with Haar and db4 wavelets, denoted as Model III and Model IV, respectively. Experimental results show that Haar and db4 wavelets perform worse in detection than db6 wavelets, mostly because db6 provides smoother basis functions and greater time-frequency localization, which make it easier to extract subtle deepfake-related clues. Subsequently, we conducted a comparative experiment on the LSTM-attention module of the backend classifier, where in Model V substituted LSTM with GRU. The findings indicate that LSTM outperforms GRU in terms of detection performance. LSTM excels in capturing long-term dependencies, resulting in superior performance on tasks that require sensitivity to long sequence features. Ultimately, the traditional RNN model was substituted with the convolutional network resnet as the backend classifier. The findings indicate that the detection performance of Model VI is inferior to that of the RNN network. This indicates that RNN-type networks are more effective than CNNs in capturing time-dependent features when processing complex speech signals.

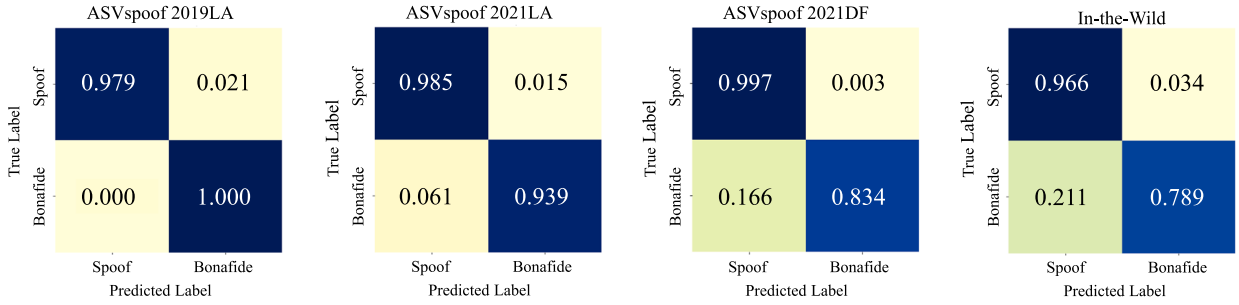
5.4. T-SNE visualization

To qualitatively assess the discriminative capability of the proposed MWAADD model, we employ t-SNE to visualize the distribution of



(a) 19LA evaluation dataset (b) 21LA evaluation dataset (c) 21DF evaluation dataset (d) ITW evaluation dataset

Fig. 5. t-SNE visualization of MWAADD on four datasets: (a) ASVspoof 2019 LA evaluation dataset, (b) ASVspoof 2021 LA dataset, (c) ASVspoof 2021 DF dataset, and (d) In-the-Wild dataset. Each plot is generated by randomly selecting 1000 samples from each class, where bonafide samples are shown in blue and spoof samples in orange.



(a) 19LA evaluation dataset (b) 21LA evaluation dataset (c) 21DF evaluation dataset (d) ITW evaluation dataset

Fig. 6. Confusion matrices of MWAADD on four datasets: (a) ASVspoof 2019 LA evaluation set, (b) ASVspoof 2021 LA dataset, (c) ASVspoof 2021 DF dataset, and (d) In-the-Wild dataset. Each confusion matrix is normalized to display the ratio of correct and incorrect predictions for each class, where the cells represent true positives, false positives, true negatives, and false negatives.

learned representations in a two-dimensional space. Fig. 5 illustrates the visualization results across four test sets: 19LA, 21LA, 21DF, and ITW(Fig. 5(a), Fig. 5(b), Fig. 5(c), Fig. 5(d)). For each dataset, we randomly sample 1000 bonafide and 1000 spoofed utterances to ensure balanced class representation. These embeddings are extracted from the penultimate layer of the MWAADD model and projected into 2D space via t-SNE with fixed parameters and random seed to ensure reproducibility and visual consistency. In the resulting plots, bonafide samples are marked in blue and spoofed samples in orange. We observe a clear separation between the two classes across all test sets, indicating that the model learns distinct embedding patterns for genuine and forged speech. Notably, spoofed samples exhibit a more scattered distribution, which may reflect the model's ability to capture variations introduced by diverse synthesis and compression techniques. In contrast, the bonafide samples are more tightly clustered, suggesting the extraction of consistent and compact representations from natural speech. Although t-SNE offers only a coarse visualization of the embedding space, it serves as a complementary analysis tool, supporting our quantitative results by providing intuitive evidence of the model's robustness and generalization. This visualization provides insight into the interpretability of the learned representations, showing that MWAADD captures distinguishable patterns between bonafide and spoofed speech.

5.5. Confusion matrices

We use the confusion matrix to evaluate the classification model's performance because it summarizes the model's predictions for each class in a detailed manner. The confusion matrix is normalized to show the proportion of correct and incorrect predictions for each class, providing insights into how well the model distinguishes between the bonafide and spoof categories. In the visualization, the matrix cells display the ratio of true positives, false positives, true negatives, and false negatives for both classes. As shown in Fig. 6, we plot the confusion matrix of MWAADD on the four test sets of 19LA, 21LA, 21DF and ITW. It is clearly found that our proposed model achieves lower false positive rate and false negatives rate in the four test machine evaluations, proving that our framework has efficient detection performance.

Table 7
Impact of the number of multi-scale attention wavelet neural operator (MAWNO) modules on detection performance (EER%).

Model	2019LA		2021LA		2021DF
	EER	mt-DCF	EER	mt-DCF	EER
L = 1	0.76	0.3563	6.92	0.3556	3.39
L = 2	0.64	0.0200	6.45	0.3556	3.96
L = 6	0.36	0.0125	6.00	0.3529	3.32
L = 8	0.46	0.0147	5.81	0.3471	3.31
L = 4	**0.18**	**0.0050**	**4.15**	**0.2900**	**2.55**

5.6. The impact of the number of mawno modules

We performed extensive ablation studies by varying L in $1, 2, 4, 6, 8$ to support the choice of the number of MAWNO blocks $L = 4$. The 19LA, 21LA, and 21DF datasets were used in these experiments, and as Table 7 illustrates, the detection performance peaked at $L = 4$. The findings imply that using too few MAWNO blocks leads to suboptimal performance due to inadequate feature representation, while increasing the number of blocks beyond 4 does not further improve detection accuracy but rather degrades performance, most likely because of increased model complexity that introduces overfitting or prevents efficient optimization.

6. Conclusion

In this study, we propose a novel multi-feature wavelet attention neural network for audio deepfake detection. We integrate wav2vec2 XLSR (XLSR) with speech emotion recognition (SER) pre-trained features to serve as the input for multi-scale wavelet attention neural network (MWANN). XLSR extracts comprehensive speech sequence features, whereas the SER module captures the absent emotion features in the manipulated speech. By combining these two feature types, we obtain more discriminative and comprehensive representations, which are then fed into MWANN for classification. MWANN comprises multi-scale attention wavelet transform layers and an LSTM-attention back-end classifier. We apply wavelet transform to audio forgery detection to enhance the classification capability through the extraction of more detailed information. Following the wavelet transform, the attention mechanism is applied to emphasize high-frequency regions that may contain forged artifacts. Finally, the LSTM-attention module uses the information extracted in the early stage to further improve the classification performance of real and fake speech. However, the performance of this method on the ASVspoof 2021LA dataset is not ideal, which may be due to the model's sensitivity to transmission noise. Future research will improve forgery detection methods targeting channel effects to enhance the robustness and adaptability of the model.

CRediT authorship contribution statement

Bo Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Rui Wang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Wei Wang:** Writing – review & editing, Methodology, Conceptualization. **Zhongjie Ba:** Writing – review & editing, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Zaynab Almutairi, Hebah Elgibreen, A review of modern audio deepfake detection methods: challenges and future directions, Algorithms 15 (5) (2022) 155.

[2] Nicholas Diakopoulos, Deborah Johnson, Anticipating and addressing the ethical implications of deepfakes in the context of elections, New Media Soc. 23 (7) (2021) 2072–2098.

[3] Cunhang Fan, Mingming Ding, Jianhua Tao, Ruibo Fu, Jiangyan Yi, Zhengqi Wen, Zhao Lv, Dual-branch knowledge distillation for noise-robust synthetic speech detection, IEEE/ACM Trans. Audio Speech Lang. Process. (2024).

[4] Yeqing Ren, Haipeng Peng, Lixiang Li, Xiaopeng Xue, Yang Lan, Yixian Yang, Generalized voice spoofing detection via integral knowledge amalgamation, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 2461–2475.

[5] Li Wang, Lingyun Yu, Yongdong Zhang, Hongtao Xie, Generalizable speech spoofing detection against silence trimming with data augmentation and multi-task meta-learning, IEEE/ACM Trans. Audio Speech Lang. Process. (2024).

[6] Li Zhang, Dezong Zhao, Chee Peng Lim, Houshyar Asadi, Haoqian Huang, Yonghong Yu, Rong Gao, Video deepfake classification using particle swarm optimization-based evolving ensemble models, Knowl.-Based Syst. 289 (2024) 111461.

[7] Catherine Stupp, Fraudsters used AI to mimic ceo's voice in unusual cybercrime case, Wall Str. J. 30 (08) (2019).

[8] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, Cuong M Nguyen, Deep learning for deepfakes creation and detection: A survey, Comput. Vis. Image Underst. 223 (2022) 103525.

[9] Zahra Khanjani, Gabrielle Watson, Vandana P. Janeja, How deep are the fakes? focusing on audio deepfake: A survey, 2021, arXiv preprint arXiv:2111.14203.

[10] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, Vrince Vimal, Deepfake generation and detection: Case study and challenges, IEEE Access (2023).

[11] Zahid Akhtar, Deepfakes generation and detection: a short survey, J. Imaging 9 (1) (2023) 18.

[12] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, Yang Liu, Countering malicious deepfakes: Survey, battleground, and horizon, Int. J. Comput. Vis. 130 (7) (2022) 1678–1734.

[13] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, Joseph K Liu, A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities, Neurocomput 513 (2022) 351–371.

[14] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, Yan Zhao, Audio deepfake detection: A survey, 2023, arXiv preprint arXiv:2308.14970.

[15] Enes Altuncu, Virginia N.L. Franqueira, Shujun Li, Deepfake: definitions, performance metrics and standards, datasets and benchmarks, and a meta-review, 2022, arXiv preprint arXiv:2208.10913.

[16] Zahid Akhtar, Thanvi Lahari Pendyala, Virinchi Sai Athmakuri, Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve, Forensic Sci. 4 (3) (2024) 289–377.

[17] Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, Ruibo Fu, An initial investigation for detecting vocoder fingerprints of fake audio, in: Proc. 1st Int. Workshop Deepfake Detect. Audio Multimedia, 2022, pp. 61–68.

[18] Nicolas M Müller, Philip Sperl, Konstantin Böttinger, Complex-valued neural networks for voice anti-spoofing, 2023, arXiv preprint arXiv:2308.11800.

[19] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, Rouba Borghol, Deepfake audio detection via MFCC features using machine learning, IEEE Access 10 (2022) 134018–134028.

[20] Zahra Khanjani, Gabrielle Watson, Vandana P. Janeja, Audio deepfakes: A survey, Front. Big Data 5 (2023) 1001063.

[21] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, Hafiz Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Appl. Intell. 53 (4) (2023) 3974–4026.

[22] Ahmed A AlSabhany, Ahmed Hussain Ali, Farida Ridzuan, AH Azni, Mohd Rosmadi Mokhtar, Digital audio steganography: Systematic review, classification, and analysis of the current state of the art, Comput. Sci. Rev. 38 (2020) 100316.

[23] Abderrahim Fathan, Jahangir Alam, Woohyun Kang, Multiresolution decomposition analysis via wavelet transforms for audio deepfake detection, in: International Conference on Speech and Computer, Springer, 2022, pp. 188–200.

[24] Bo Wang, Yeling Tang, Fei Wei, Zhongjie Ba, Kui Ren, FTDKD: Frequency-time domain knowledge distillation for low-quality compressed audio deepfake detection, IEEE/ACM Trans. Audio Speech Lang. Process. (2024).

[25] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al., XLS-r: Self-supervised cross-lingual speech representation learning at scale, 2021, arXiv preprint arXiv:2111.09296.

[26] Yi Zhu, Saurabh Powar, Tiago H. Falk, Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Workshops., IEEE, 2024, pp. 139–143.

[27] Junyan Wu, Qilin Yin, Ziqi Sheng, Wei Lu, Jiwu Huang, Bin Li, Audio multi-view spoofing detection framework based on audio-text-emotion correlations, IEEE Trans. Inf. Forensics Secur. (2024).

[28] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, Stefano Tubaro, Deepfake speech detection through emotion recognition: a semantic approach, in: ICASSP 2022-2022 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2022, pp. 8962–8966.

[29] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 2823–2832.

[30] Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, Vadim Shchemelinin, STC anti-spoofing systems for the asvspoof 2015 challenge, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2016, pp. 5475–5479.

[31] Priyanka Gupta, Hemant A. Patil, Morse wavelet transform-based features for voice liveness detection, Comput. Speech & Lang. 84 (2024) 101571.

[32] Utsav Avaiya, Naman Badlani, Rajas Baadkar, Kiran Talele, Audio deepfake detection using deep scattering network, in: 2024 9th International Conference on Communication and Electronics Systems, ICCES, IEEE, 2024, pp. 2063–2069.

[33] Jiayang Su, Junbo Ma, Songyang Tong, Enze Xu, Minghan Chen, Multiscale attention wavelet neural operator for capturing steep trajectories in biochemical systems, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (13) 2024, pp. 15100–15107.
[34] Rui Wang, Zirui Chen, Bo Wang, Zhongjie Ba, Kui Ren, AWaveFormer: Audio wavelet transformer network for generalized audio deepfake detection, IEEE Trans. Audio Speech Lang. Process. (2025).
[35] Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, Zhengqi Wen, Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 11736–11745.
[36] Yujie Chen, Jiangyan Yi, Cunhang Fan, Jianhua Tao, Yong Ren, Siding Zeng, Chu Yuan Zhang, Xinrui Yan, Hao Gu, Jun Xue, et al., Region-based optimization in continual learning for audio deepfake detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, (22) 2025, pp. 23651–23659.
[37] Ashutosh Anshul, Eng Siong Chng, Deepu Rajan, Av representation learning via audio shift prediction for multimodal deepfake detection and temporal localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2026, pp. 2553–2563.
[38] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, Anthony Larcher, End-to-end anti-spoofing with rawnet2, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process, IEEE, 2021, pp. 6369–6373.
[39] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, Nicholas Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process, IEEE, 2022, pp. 6367–6371.
[40] Cunhang Fan, Jun Xue, Jianhua Tao, Jiangyan Yi, Chenglong Wang, Chengshi Zheng, Zhao Lv, Spatial reconstructed local attention Res2Net with F0 subband for fake speech detection, Neural Netw. 175 (2024) 106320.
[41] Xin Wang, Junichi Yamagishi, Investigating self-supervised front ends for speech spoofing countermeasures, 2021, arXiv preprint arXiv:2111.07725.
[42] Yang Xiao, Rohan Kumar Das, XLSR-mamba: A dual-column bidirectional state space model for spoofing attack detection, IEEE Signal Process. Lett. (2025).
[43] Eros Rosello, Alejandro Gómez Alanís, Angel M Gomez, Antonio M Peinado, N Harte, J Carson-Berndsen, G Jones, A conformer-based classifier for variable-length utterance processing in anti-spoofing, in: Interspeech, vol. 2023, 2023, pp. 5281–5285.
[44] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, Junichi Yamagishi, The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2022) 813–825.
[45] Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, Yunhe Wang, A robust audio deepfake detection system via multi-view feature, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 13131–13135.
[46] Haibin Wu, Xu Li, Andy T Liu, Zhiyong Wu, Helen Meng, Hung-Yi Lee, Improving the adversarial robustness for speaker verification by self-supervised learning, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2021) 202–217.
[47] Soumya Dutta, Sriram Ganapathy, Leveraging content and acoustic representations for efficient speech emotion recognition, 2024, pp. arXiv–2409, ArXiv E-Prints.
[48] Yuanchao Li, Yumnah Mohamied, Peter Bell, Catherine Lai, Exploration of a self-supervised speech model: A study on emotional corpora, in: 2022 IEEE Spoken Language Technology Workshop, SLT, IEEE, 2023, pp. 868–875.
[49] Jiayang Su, Junbo Ma, Songyang Tong, Enze Xu, Minghan Chen, Multiscale attention wavelet neural operator for capturing steep trajectories in biochemical systems, in: Proc. AAAI Conf. Artif. Intell., vol. 38, (13) 2024, pp. 15100–15107.
[50] Lukas P.A. Arts, Egon L. van den Broek, The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis, Nat. Comput. Sci. 2 (1) (2022) 47–58.
[51] Ruyue Liu, Rong Yin, Yong Liu, Weiping Wang, ASWT-sgnn: Adaptive spectral wavelet transform-based self-supervised graph neural network, in: Proc. AAAI Conf. Artif. Intell., vol. 38, (12) 2024, pp. 13990–13998.
[52] Pranav Jeevan, Amit Sethi, Wavemix: resource-efficient token mixing for images, 2022, arXiv preprint arXiv:2203.03689.
[53] Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, Amit Sethi, Wavemixsr: Resource-efficient neural network for image super-resolution, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5884–5892.
[54] Mingyi Chen, Xuanji He, Jing Yang, Han Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process. Lett. 25 (10) (2018) 1440–1444.
[55] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, Shrikanth S Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[56] Zhen Zhang, Yicheng Ye, Binyu Luo, Guan Chen, Meng Wu, Investigation of microseismic signal denoising using an improved wavelet adaptive thresholding method, Sci. Rep. 12 (1) (2022) 22186.
[57] Xiao Xu, Yang Wang, Xinru Wei, Fei Wang, Xizhe Zhang, Attention-based acoustic feature fusion network for depression detection, Neurocomput. 601 (2024) 128209.
[58] Yidi Li, Hong Liu, Hao Tang, Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking, in: Proc. AAAI Conf. Artif. Intell., vol. 36, (2) 2022, pp. 1456–1463.
[59] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, Yuejie Zhang, Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning, in: Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 3884–3892.
[60] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech, IEEE Trans. Biom. Behav. Identity Sci. 3 (2) (2021) 252–265.
[61] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 2507–2522.
[62] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, Konstantin Böttinger, Does audio deepfake detection generalize?, 2022, arXiv preprint arXiv:2203.16263.
[63] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.(2016), 2016, http://datashare.is.ed.ac.uk/handle/10283/2651.
[64] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, Zhenhua Ling, The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods, 2018, arXiv preprint arXiv:1804.04262.
[65] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, Tomoki Toda, Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion, 2020, arXiv preprint arXiv:2008.12527.
[66] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, 2019, arXiv preprint arXiv:1904.05441.
[67] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, Nicholas Evans, Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, 2022, arXiv preprint arXiv:2202.12233.
[68] Guang Hua, Andrew Beng Jin Teoh, Haijian Zhang, Towards end-to-end synthetic speech detection, IEEE Signal Process. Lett. 28 (2021) 1265–1269.