


RESEARCH

Open Access



Shallow and deep feature fusion for digital audio tampering detection

Zhifeng Wang^{1*} , Yao Yang², Chunyan Zeng^{2*}, Shuai Kong², Shixiong Feng² and Nan Zhao²

*Correspondence:
zfwang@ccnu.edu.cn;
cyzeng@hbut.edu.cn

¹Department of Digital Media Technology, Central China Normal University, Luoyu Road 152, Wuhan 430079, China

²Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Nanli Road 28, Wuhan 430068, China

Abstract

Digital audio tampering detection can be used to verify the authenticity of digital audio. However, most current methods use standard electronic network frequency (ENF) databases for visual comparison analysis of ENF continuity of digital audio or perform feature extraction for classification by machine learning methods. ENF databases are usually tricky to obtain, visual methods have weak feature representation, and machine learning methods have more information loss in features, resulting in low detection accuracy. This paper proposes a fusion method of shallow and deep features to fully use ENF information by exploiting the complementary nature of features at different levels to more accurately describe the changes in inconsistency produced by tampering operations to raw digital audio. Firstly, the audio signal is band-pass filtered to obtain the ENF component. Then, the discrete Fourier transform (DFT) and Hilbert transform are performed to obtain the phase and instantaneous frequency of the ENF component. Secondly, the mean value of the sequence variation is used as the shallow feature; the feature matrix obtained by framing and reshaping of the ENF sequence is used as the input of the convolutional neural network; the characteristics of the fitted coefficients are obtained by curve fitting. Then, the local details of ENF are obtained from the feature matrix by the convolutional neural network, and the global information of ENF is obtained by fitting coefficient features through deep neural network (DNN). The depth features of ENF are composed of ENF global information and local information together. The shallow and deep features are fused using an attention mechanism to give greater weights to features useful for classification and suppress invalid features. Finally, the tampered audio is detected by downscaling and fitting with a DNN containing two fully connected layers, and classification is performed using a Softmax layer. The method achieves 97.03% accuracy on three classic databases: Carioca 1, Carioca 2, and New Spanish. In addition, we have achieved an accuracy of 88.31% on the newly constructed database GAUDI-DI. Experimental results show that the proposed method is superior to the state-of-the-art method.

Keywords: Electronic network frequency, Audio forensics, Deep learning, Feature fusion

1 Introduction

More and more software for digital audio editing has been developed recently, and it has become much easier to edit, tamper and forge digital audio. The audio editing software (such as Adobe Audition, WavePad, and Ocenaudio) makes it easy for ordinary people

to delete, insert, copy and paste digital audio tampering, resulting in changes in audio semantics. Moreover, with the continuous enhancement of audio editing technology, it is impossible to tell whether the audio has been tampered with by the human ear. However, some edited digital audio may be misused, especially in essential security applications such as courts, politics, or business, which may cause serious consequences [1]. For digital audio that is deliberately or even maliciously tampered with, it is essential to develop efficient digital audio tampering detection methods.

Digital audio tampering detection methods include active detection and passive detection [2]. There are two standard technologies in active detection methods: digital watermark and digital signature of digital audio. These two technologies require a pre-embedding watermark or signature in the audio to be detected. However, most of the audio to be detected is not pre-embedded with this additional information in practice. The passive detection methods are more practical to use the characteristics of digital audio itself to detect tampering without adding any additional information [3].

At present, passive audio tamper detection methods are mainly based on visual contrast analysis of frequency continuity of electronic network based on digital audio and standard ENF database, or feature extraction of ENF signals and classification by machine learning method. ENF database is usually challenging to obtain, and the feature expression of the visualization method is weak. In contrast, the feature information loss of the machine learning method is significant, resulting in low detection accuracy. This paper proposes an audio tamper detection method based on the fusion of shallow and deep features to solve this problem. Firstly, the audio signal is bandpass filtered to obtain the ENF component. Then, the phase and instantaneous frequency of the ENF component are obtained by DFT and Hilbert transform. Second, the ENF phase and instantaneous frequency are processed in three ways. The mean value of sequence variation is taken as the shallow feature. The feature matrix obtained by framing and reshaping of the ENF sequence is the input to the convolutional neural network. Curve fitting was carried out to obtain the characteristics of fitting coefficients. The problem of the unequal length of ENF features is proposed to be solved by framing and fitting processing methods to make them suitable for the input of neural networks. Then, in the neural network, the feature matrix is input into the convolutional neural network to obtain the local details of ENF, and the fitting coefficient feature obtains the global information of ENF through DNN. The global information and local details together constitute the deep features of ENF. The characteristics of the ENF phase and frequency features are fully considered to obtain deep features containing both global and local information about ENF. The attention mechanism fuses the shallow and deep features. The fusion of shallow and deep features exploits the complementary nature of features at different levels to more accurately describe the changes in inconsistency produced by tampering operations to natural digital audio. Finally, the DNN classifier with two fully connected layers was used to fit, and Softmax was used to classify and detect tampered audio. The main contributions of this paper are as follows:

1. We propose a novel shallow and deep feature fusion-based framework for digital audio tampering detection by automatically analysing the continuity of ENF. With using deep learning methods to learn features for tampering detection automati-

cally, the algorithm has a higher degree of automation than threshold and visualization methods. In addition, the proposed framework achieves state-of-the-art performance on three publicly available dataset Carioca 1, Carioca 2, and New Spanish.

2. On the one hand, through the fusion of shallow and deep features, it will acquire the complementarity of different features, which is a more comprehensive description of audio ENF features and can be used to improve algorithm robustness and model generalization capabilities. On the other hand, the local and global information is obtained from the audio ENF through automatic learning to reduce information loss and further improve detection accuracy.
3. The attention mechanism is used to fuse phase and frequency features to obtain useful detailed information for tampering detection and classification task from audio through automatic learning to improve classification accuracy and model generalization ability.

The rest of this paper is organised as follows. Section 2 presents the relevant existing works in the literature. Section 3 describes the audio tampering detection framework. Section 4 presents the proposed audio tampering detection method based on shallow and deep feature fusion. The results of experiments and analyses are shown in Sect. 5. Lastly, we come to a conclusion and list some future work in Sect. 6.

2 Related work

Digital audio passive forensics realizes tampering detection by extracting and analyzing audio features. These features can be divided into traditional shallow features and deep features generated by deep neural networks.

2.1 Detection methods based on shallow features

The features contained in digital audio are divided into three categories, which are (1) environment and device characteristics; (2) time-domain and frequency domain characteristics; and (3) electronic network frequency characteristics.

(1) *Environment and device characteristics in audio* Digital audio is obtained by recording equipment in a particular environment, which will lead to the existence of some equipment and environment information in the audio. An audio file is regarded as edited one when there is different background information involved in the audio [4]. Malik [5] carries out endpoint detection of speech signals, extracts the attenuating signal part at the end, and uses statistical methods to model and estimate the reverberation and background noise in the attenuating signal, which is used to classify different signals. In [6], the method is tested and improved on this basis, and the robustness of the original method against MP3 compression is improved. The device information in the audio can be analyzed to determine whether the audio has been edited [7]. Cuccovillo et al. [8] analyze the microphones of recorded audio to detect the presence of multiple microphones in single audio for tampering detection. When recording audio, the surrounding environment mainly contains background noise, which can be used for audio tampering detection by analyzing background noise. In [9], according to the significant differences in the audio background noise levels of different recording environments, the similarity

of each syllable's background noise variance is compared to judge whether there is a heterogeneous splicing tampering operation in the audio.

(2) *Time-domain and frequency domain characteristics of audio* When editing digital audio, some features of the audio will be affected, resulting in abnormal changes in features, which will make the discontinuity or correlation between adjacent frames weakened. Audio tampering detection can be realized by analyzing audio time-domain and frequency domain characteristics [10]. Time-domain features are used for tampering detection as follows: Yan et al. [11] detect the smoothing processing after audio tampering through the local variance of differential signals. Yan et al. [12] took pitch sequence and formant sequence as the features of voiced fragments and realized copy-move tampering detection and location by comparing their similarity. The use of frequency domain features for tampering detection includes: Chen et al. [13] used wavelet packet singularity analysis to detect the insertion, deletion, replacement, and concatenation operations according to the singularity points generated by the weakened signal correlation caused by audio tampering. In [14], Lin et al. used the short-time Fourier transform (STFT) to reconstruct the spectral phase to offset the influence of noise and propose a feature based on the spectral phase residual and spectral phase correlation between two adjacent clear segments, so as to realize tampering detection and location at the high noise level. Xie et al. [15] combined the four characteristics of the Gammatone feature, Mel-frequency cepstral coefficients (MFCCs) feature, pitch feature, and DFT coefficients and adopted the decision tree method to conduct copy-move tampering detection.

(3) *Electronic network frequency in audio* ENF is widely used in audio forensics [16]. ENF is the transmission frequency with a nominal value of 50 or 60 Hz in the power grid. When recording audio in the electrical activity area, the ENF signal will be embedded into the audio [17]. The fluctuation of ENF in a specific area is stable and unique within a certain period [18], so ENF can be used to detect audio tampering [19, 20]. Two existing methods for audio tampering detection using ENF include database comparison and consistency analysis. Audio tampering detection can be carried out by comparing ENF in audio with the ENF database. Hua et al. [21] detect insertion, deletion, and stitching operations through the Absolute-Error-Map between the ENF signal in audio and the database. However, it is difficult to obtain the ENF database, and many studies have used ENF discontinuity to detect audio. Most tampering operations cause the ENF to change suddenly at the tampering point. Esquef et al. [22] use Hilbert transform to calculate the instantaneous frequency of ENF and propose two-pass split window (TPSW) method to estimate the change degree ENF background to achieve tampering detection. Rodriguez et al. [23, 24] detect audio tampering by extracting ENF signals and detecting ENF phase changes' consistency.

2.2 Detection methods based on deep features

With the development of deep learning and artificial intelligence technology, some scholars also use deep learning methods to deal with audio forensics tasks. Deep learning-based methods perform audio tampering detection tasks by training a deep neural network model with dataset in advance. Using a large amount of data to train the model can reduce the practical problems caused by artificially setting the threshold. The deep learning method supports higher-dimensional input features.

Combining multiple parameters in the deep neural network can better fit the audio features, learn the difference between original audio and tampered audio, and make the detection more accurate and more robust.

Tamper detection based on deep learning methods can be divided into three sub-categories. (1) Frequency domain features are used to identify audio post-processing operations. Wang et al. [25] used the features of audio after STFT transformation as the input of the convolutional neural network (CNN) to identify the post-processing operation of audio pitch transformation. (2) ENF is applied for audio recapture detection. Lin et al. [26] take ENF spectrogram as the convolutional neural network input for audio recapture detection. (3) Use the spectrogram to detect insertion and tampering in audio. Jadhav et al. [27] directly input the audio spectrum map into the convolutional neural network to detect the audio insertion tampering.

3 Audio tampering detection framework

The audio tampering detection framework is shown in Fig. 1. The tampering detection methods can be classified into active detection methods and passive detection methods. The active detection method is to embed a digital watermark, and signature in the audio when the audio is generated. When the audio is edited and tampered with, the watermark information embedded in the audio in advance will change so that the edited audio can be accurately distinguished. However, there are often no such watermarks in audio. The passive detection method uses the audio content itself as a feature, detects tampering in these features through a threshold, or trains a model through machine learning and other methods to perform tampering detection.

In the audio tampering detection task, the audio signal can be formulated by

$$y(n) = s(n) + v(n) + f(n) \quad (1)$$

where $s(n)$ represents speech content, $v(n)$ represents background noise, and $f(n)$ represents ENF. In traditional digital audio tampering detection, one of the speech contents, background noise or ENF in audio, is usually extracted and analyzed. The audio is windowed and divided into frames. Extract the time domain or frequency domain features of each frame of audio, such as pitch, reverberation, background noise, MFCC, ENF, and other time and frequency domain features. Then set the corresponding threshold to detect the abrupt changes between frames or detect it through the support vector machine (SVM) [16].

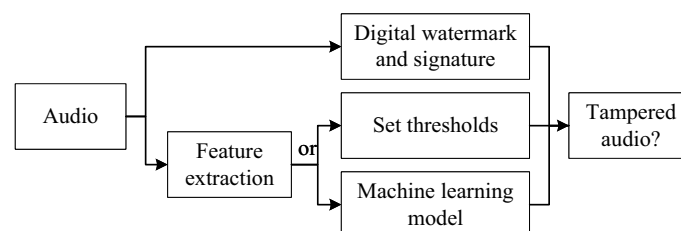


Fig. 1 Audio tampering detection framework

4 Audio tampering detection based on shallow and deep feature fusion

The audio tampering detection method based on the fusion of shallow and deep features proposed in this paper consists of three parts (Fig. 2):

1. *Phase and frequency feature extraction* First, down-sampling and band-pass are employed to filter the audio to obtain the ENF component. Then, windowing and framing process is implemented on the ENF components. Finally, DFT is used to obtain the phase feature, and Hilbert transform is applied to obtain the instantaneous frequency.
2. *Feature process* In this part, the average of ENF phase and frequency variations is calculated as the shallow features. Meanwhile, the feature matrix is obtained by framing and reshaping operations on the audio (see Sect. 4.2.2 for details). The feature matrix will be used as the convolutional neural network’s input to obtain more local information. The fit coefficients are obtained by fitting the phase to the instantaneous frequency through Sums of Sines functions [16], and the fit coefficients are the input to the DNN to give some global information compensation to the deeper features.
3. *Deep neural network* In this part, the feature matrix and fitting coefficients are input to the neural network, and the output is stitched to obtain the deep features that contain both global and local information. Finally, the deep, deep phase, and instantaneous frequency features are fused with features using the attention mechanism to

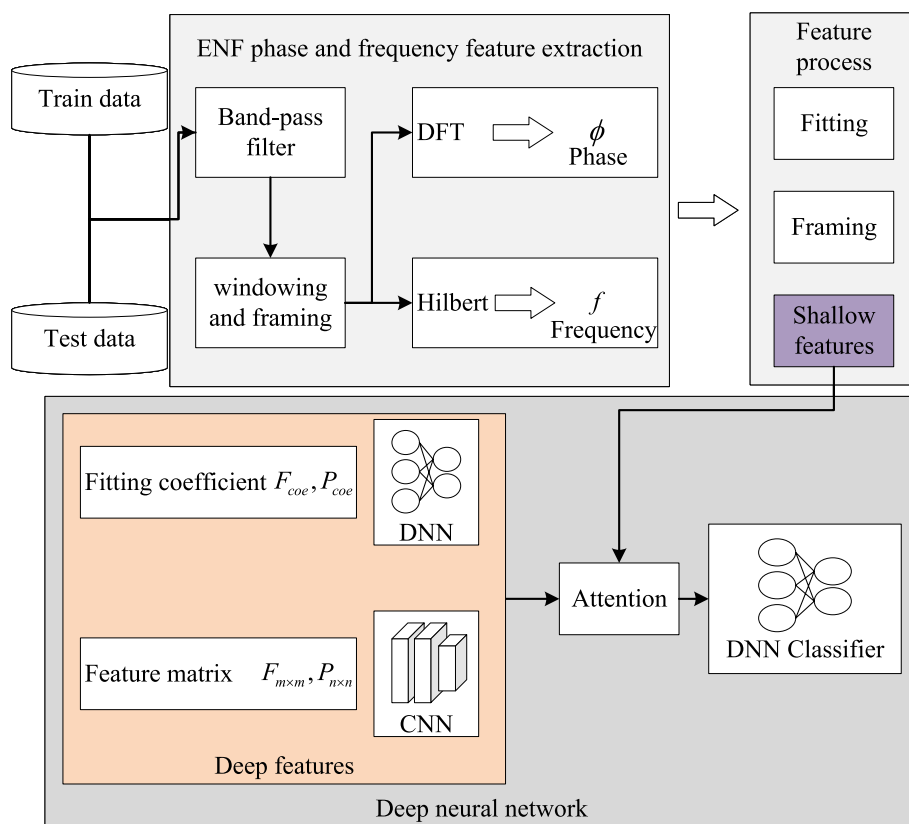


Fig. 2 Shallow and deep feature fusion method

give different weights to each feature vector value to achieve feature selection. Finally, a DNN classifier is proposed to classify the tampered audio with the authentic audio.

The specific details will be introduced later in this section.

4.1 ENF phase and frequency feature extraction

According to the method in the literature [22, 23], to obtain the phase and instantaneous frequency of ENF, we performed subsampling and bandpass filtering on the audio. Firstly, the subsampling frequency is set as 1000 Hz and 1200 Hz according to the ENF nominal frequency of 50 or 60 Hz. The purpose of this is to ensure the accuracy of ENF while reducing the amount of calculation. Then, after subsampling, bandpass filtering is carried out to obtain ENF components in the audio. We use a linear zero-phase FIR filter of order 10000 to carry out narrowband filtering. The centre frequency is ENF standard (50 Hz or 60 Hz), the bandwidth is 0.6 Hz, the passband ripple is 0.5 dB, and the stopband attenuation is 100 dB. Finally, we can obtain the phase and instantaneous frequency of ENF through DFT and Hilbert transformation.

4.1.1 DFT transformation gets the phase

The phase of ENF is obtained by discrete Fourier transform, and the phase of DFT^0 and DFT^1 is calculated. DFT^k represents the k derivative of the *DFT* transform of a signal, and DFT^0 represents the conventional DFT transform [23].

First, the approximate first derivative $x'_{ENFC}[n]$ of ENF signal $X_{ENFC}[n]$ at point n is calculated

$$x'_{ENFC}[n] = f_d(X_{ENFC}[n] - X_{ENFC}[n - 1]) \quad (2)$$

where $f_d(*)$ represents the approximate derivative operation, and $X_{ENFC}[n]$ represents the n -th point of the ENF component.

Then, Hanning window $w(n)$ was used to frame and window $x'_{ENFC}[n]$. The frame length was 10 standard ENF frequency cycles ($\frac{10}{50}$ or $\frac{10}{60}$), and the frame was moved to 1 standard ENF frequency cycle ($\frac{1}{50}$ or $\frac{1}{60}$).

$$x'_N[n] = x'_{ENFC}[n]w(n) \quad (3)$$

where $x'_N[n]$ represents the ENF signal after window addition, and $w(n)$ represents the Hanning window.

To obtain the phase ϕ_{DFT^0} of ENF and the phase ϕ_{DFT^1} of the first derivative of ENF, n -point DFT should be executed for each frame signal $x'_N[n]$ and $X_{ENFC}[n]$, respectively, to obtain $X'(k)$ and $X(k)$. Estimated frequency f_{DFT^1} based on the integer index k_{peak} of $|X'(k)|$ peak points

$$f_{DFT^1} = \frac{1}{2\pi} \frac{DFT^1[k_{peak}]}{DFT^0[k_{peak}]} \quad (4)$$

where $DFT^0[k_{peak}] = X(k_{peak})$, $DFT^1[k_{peak}] = F(k_{peak})|X'(k_{peak})|$ and $F(k_{peak})$ are scale coefficients.

$$F(k) = \frac{\pi k}{N_{\text{DFT}} \sin\left(\frac{\pi k}{N_{\text{DFT}}}\right)} \quad (5)$$

where N_{DFT} represents the number of discrete Fourier transform points, and k is the index of peak point.

Now the ENF phase ϕ_{DFT^0} of the conventional DFT transformation can be calculated, $\phi_{\text{DFT}^0} = \arg[X(k_{\text{peak}})]$. Through Eq. (6), ϕ_{DFT^1} [23] can be calculated.

$$\begin{cases} \phi_{\text{DFT}^1} = \arctan\left\{\frac{\tan(\theta)[1-\cos(\omega_0)]+\sin(\omega_0)}{1-\cos(\omega_0)-\tan(\theta)\sin(\omega_0)}\right\} \\ \theta \approx (k_{\text{DFT}^1} - k_{\text{low}}) \frac{\theta_{\text{high}} - \theta_{\text{low}}}{k_{\text{high}} - k_{\text{low}}} + \theta_{\text{low}} \end{cases} \quad (6)$$

where $\omega_0 \approx 2\pi f_{\text{DFT}^1}/f_d$, f_d are heavy sampling frequency, $k_{\text{DFT}^1} = f_{\text{DFT}^1} N_{\text{DFT}}/f_d$, $k_{\text{low}} = \text{floor}[k_{\text{DFT}^1}]$, $k_{\text{high}} = \text{ceil}[k_{\text{DFT}^1}]$, $\text{floor}[a]$ is the maximum integer less than a , and $\text{ceil}[b]$ is the minimum integer greater than b . Since the calculated ϕ_{DFT^1} has two possible values, ϕ_{DFT^0} is used as a reference, and the value closest to ϕ_{DFT^0} in ϕ_{DFT^1} is selected as the final ϕ_{DFT^1} .

4.1.2 The Hilbert transform captures the instantaneous frequency

Hilbert transformation [22] was performed on the filtered ENF signal $X_{\text{ENFC}}[n]$ to obtain the ENF instantaneous frequency $f[n]$. So first, we get the analytic function of $X_{\text{ENFC}}[n]$

$$x^{(a)}_{\text{ENFC}}[n] = X_{\text{ENFC}}[n] + i * H\{X_{\text{ENFC}}[n]\} \quad (7)$$

where $H\{*\}$ stands for Hilbert transformation, $i = \sqrt{-1}$. Instantaneous frequency $f[n]$ is the rate of change of $H\{X_{\text{ENFC}}[n]\}$ phase angle.

The parasitic oscillation generated by the numerical approximation during the Hilbert transformation needs to be removed after the instantaneous frequency $f[n]$ obtained. The fifth-order elliptic IIR filter was used to carry out the low-pass filter on $f[n]$ to remove oscillation. The filter's central frequency is ENF standard frequency, the bandwidth is 20 Hz, the passband ripple is 0.5 dB, and the stopband attenuation is 64 dB. Due to the boundary effect of frequency estimation, the head and tail of $f[n]$ are removed for about 1 s. Finally, f_{hil} of instantaneous frequency estimation of ENF component is obtained.

4.2 Shallow feature acquisition and deep feature preparation

We use the average of ENF phase and instantaneous frequency changes as shallow features. To obtain the deep features, we use a convolutional neural network better to learn the details of ENF phases and instantaneous frequencies. We frame, reshape and fit the ENF phase and frequency to get the input to the neural network and feed it to the neural network to obtain the depth features for the training phase of the network.

4.2.1 Acquire shallow features

The estimated phase ϕ_{DFT^0} , ϕ_{DFT^1} and Hilbert instantaneous frequency f_{hil} are put into Eq. (8) to obtain the statistical feature $F_{01f} = [F_0, F_1, F_f]$, which reflects the abrupt transition of ENF phase and instantaneous frequency [19].

$$\begin{cases} F_{0,1} = 100 \log \left\{ \frac{1}{N_{\text{Block}}-1} \sum_{n_b=2}^{N_{\text{Block}}} [\hat{\phi}'(n_b) - m_{\hat{\phi}'}]^2 \right\} \\ F_f = 100 \log \left\{ \frac{1}{\text{len}-1} \sum_{n=2}^{\text{len}} [f'(n) - m_{f'}]^2 \right\} \end{cases} \quad (8)$$

where $\hat{\phi}'(n_b) = \hat{\phi}(n_b) - \hat{\phi}(n_b - 1)$, $2 \leq n_b \leq N_{\text{Block}}$. $\hat{\phi}(n_b)$ is the estimated phase of the corresponding n_b frame. $m_{\hat{\phi}'}$ represents the average value of $\hat{\phi}'(n_b)$ from $n_b = 2$ to N_{Block} . $\text{len} = \text{length}(X_{\text{ENFC}}[n])$, $f'(n) = f(n) - f(n - 1)$. $f(n)$ is the instantaneous frequency of the n th sampling point, and $m_{f'}$ represents the average value of $f'(n)$ from $n = 2$ to len .

4.2.2 Obtaining the input of deep features $F_{m \times m}, P_{n \times n}$

The deep features proposed in this paper consist of two parts, firstly, the local detail information obtained by the feature matrix $F_{m \times m}$ and $P_{n \times n}$ through the convolutional neural network, obtained by framing and reshaping operations. The second is the global information obtained by fitting coefficients through DNN. Finally, the global information is stitched with detailed information to obtain deep features.

To reduce information loss, we acquire the deep features by convolutional neural networks. Therefore, we designed a framing approach for obtaining the input of the convolutional neural network so that the audio ENF phase or frequency of unequal lengths through the dataset becomes a matrix of $m \times m$. Where m is the frame length (the audio determines the frame length with the longest duration in the data), and each row in the matrix is one frame, and the frame shift s of each audio is computed adaptively. The detailed steps are listed in following Algorithm 1.

Algorithm 1 Obtaining the input of deep features $F_{m \times m}, P_{n \times n}$

- 1: dataset: The audio length is about 9-35s
 - 2: **for** All audio data **do**
 - 3: Get the maximum length of audio
 - 4: DFT and Hilbert transform for the longest duration audio
 - 5: Get the maximum ϕ length $\text{len}(\phi)_{\max}$ and f' length $\text{len}(f)_{\max}$
 - 6: Calculates the frame length $m(n) = \text{ceil}(\sqrt{X_{\max}})$, $X_{\max} = \text{len}(\phi)_{\max}, \text{len}(f)_{\max}$
 - 7: **for** All audio data **do**
 - 8: DFT and Hilbert transform
 - 9: Calculate the overlap and divide the frame. $\text{overlap} = m(n) - \text{ceil}\left(\frac{X - m(n)}{m(n)-1}\right)$ Reshape into feature matrix $F_{m \times m}, P_{n \times n}$
 - 10: **return** $F_{m \times m}, P_{n \times n}$
-

4.2.3 Curve fitting for fitting coefficient

We performed a reshape operation when obtaining the feature matrix of the convolutional neural network input, which may result in the loss of global information of the sequence, so we fit the ENF phase and frequency sequences and used the fit coefficients as compensation for the global information. The ENF phase and instantaneous frequency are curve-fitted to extract the fit coefficients containing the global information. We use the MATLAB fitting toolbox to extract the fitting coefficients using six Sum of Sines functions to fit the phase, and frequency features $F_{\text{coe}}, P_{\text{coe}} = [a_1, b_1, c_1, \dots, a_j, b_j, c_j]$ ($0 < j \leq 6$). The Sum of Sines functions

$$y = \sum_{j=1}^6 a_j \sin(b_j x + c_j) \quad (9)$$

4.3 Shallow and deep feature fusion network

There is information loss by only going through shallow features, resulting in the inability to obtain higher detection accuracy and model generalization. The duration of each detected audio is different, so the obtained phase feature length and frequency feature length are also different. As shown in Fig. 3, in the tampering detection method based on the fusion of shallow and deep features proposed in this paper, the phase and instantaneous frequency features of the ENF are first processed to make them suitable for automatic learning of the neural network and reduce information loss. Then, the depth features of ENF are obtained by the neural network to understand better the difference between tampered audio and real audio by automatic learning. Then, feature fusion is performed using attention, and finally, the detection results are output.

4.3.1 Neural networks of deep features

As shown in Fig. 3, the shallow feature F_{123} , which are extracted through the framing and the Sum of Sines fitting, reflects the sudden change of ENF phase and frequency, but its statistical feature is only a single value, and detailed information about ENF phase and frequency will be lost. When it is only used for audio tampering detection, it may cause misjudgement due to the insignificant fluctuation of ENF in the tampered area, or the interference of low-frequency noise on ENF. In order to reduce the occurrence of this situation, we use the convolutional neural network to obtain ENF detailed information as deep features and use the attention mechanism to combine deep features with shallow features to reduce misjudgements and improve the generalization ability of the model.

The deep features proposed in this paper are obtained from the fitting coefficients and the feature matrix. The fitting coefficients are passed through two fully connected layers with 32 neurons to obtain the ENF phase and global frequency information. A convolutional neural network extracts the phase and frequency feature matrices to obtain detailed information about the ENF phase and frequency. The size of the phase feature matrix $P_{n \times n}$ is different from that of the frequency feature matrix $F_{m \times m}$. As the size of the feature matrix is $n \times n$, $m \times m$, the frame length is determined by the longest audio in the audio data, and the longest duration of the digital audio that this network can detect is 35 s. Since the longest audio in our dataset is 35 s, the length of the phase and instantaneous frequency sequences obtained by DFT and Hilbert transform are 2055 and 37,281, so our frame length in the deep feature is set to 46,194 by the steps in 1. The number of convolution blocks for phase features is 2, and for instantaneous frequency, convolution blocks are 3. When the longest length of the audio to be measured increases or decreases, the number of convolution blocks should be increased or decreased as appropriate.

We use two convolution blocks to extract features from the phase feature matrix $P_{n \times n}$ and three convolution blocks to extract features from the frequency feature matrix

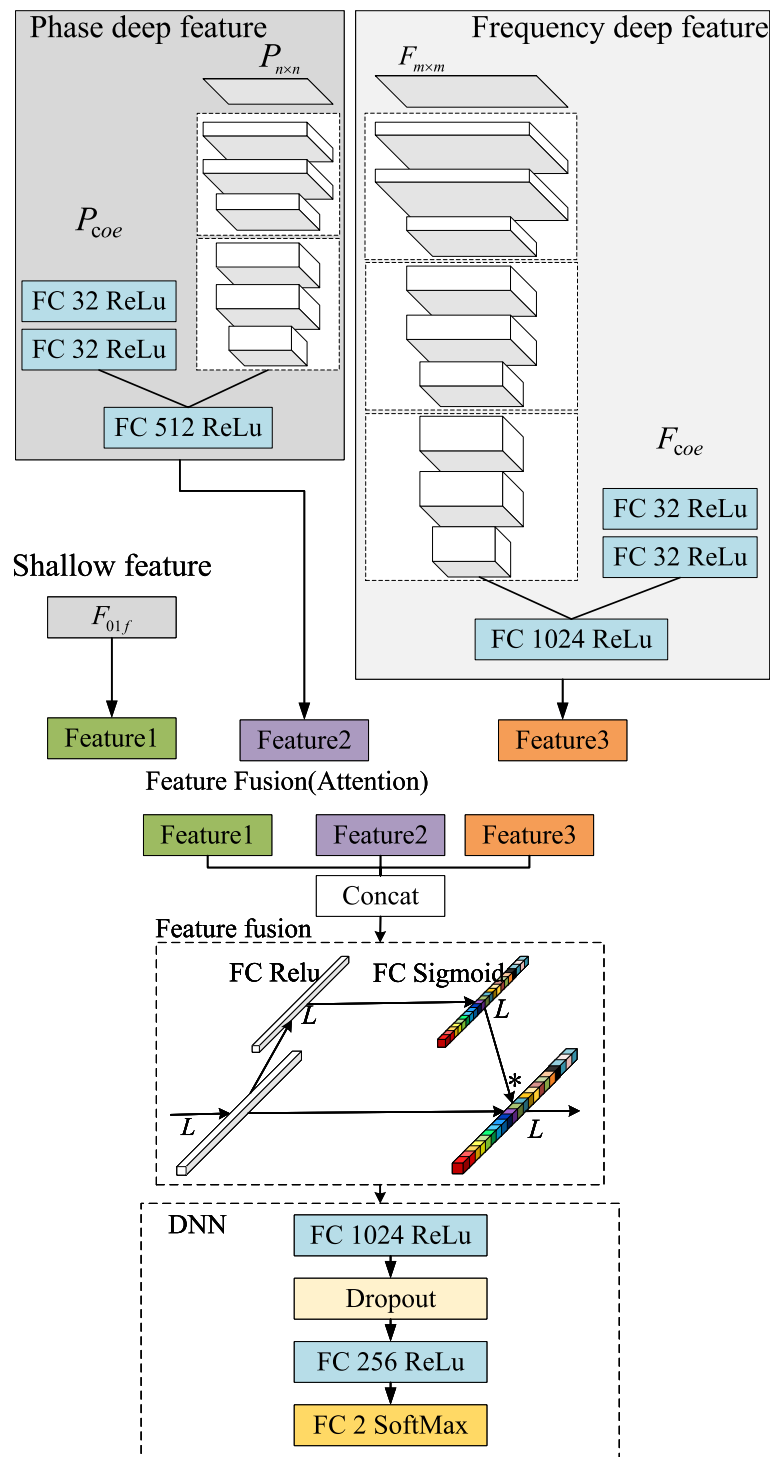


Fig. 3 Shallow and deep feature fusion network

$F_{m \times m}$. Each convolutional block consists of two identical convolutional layers with one pooling layer (the number of filters for the three convolutional blocks is 32,64,128. The convolutional kernel size is $3 * 3$ and the step is 1. The Maxpooling layer pool size is 3). Detailed information of the ENF phase and frequency sequence can be obtained by

using the local sensing property of the convolutional neural network. The pooling layer is used for dimensionality reduction to reduce the number of parameters, avoid overfitting, and improve the model's fault tolerance and generalization ability. Also, because the convolutional neural network has fewer parameters, it can obtain better classification results with less training time.

Frequency fitting coefficient F_{coe} , two fully connected layers were used to fit its characteristics. (The number of neurons was 32, 32, and the activation function was Relu.) The output of the convolution block is dimensioned through a layer of fully connected with 1024 neurons, then splicing with the fitting coefficient features after DNN fitting. Finally, the deep feature is obtained through the fully connected layer of 1024 neurons. The deep feature contains both the global information of the fitting coefficient and the local information obtained by the convolutional neural network.

4.3.2 The attention mechanism of feature fusion

We use the attention mechanism [28] to fuse shallow and deep features. In the feature fusion part (as shown in Fig. 3), firstly, we concatenate the shallow and deep features of phase and frequency to obtain the input of length L . Then, to get the weight of each feature, we will input the fully connected layer through the two activation functions for ReLU and Sigmoid. We use the ReLU activation function to enhance the nonlinearity and obtain the weight through Sigmoid. Finally, the input features are multiplied by the weights.

The attention fusion mechanism used in this paper uses the Sigmoid activation function instead of Softmax to obtain the weights because the primary purpose of the attention mechanism used in this paper is to suppress invalid features, not to find the optimal features. There is no need for each feature value in the shallow and deep layers to compete for weights. This is because the primary purpose of the attention mechanism used in this paper is to suppress the invalid features, not to find the optimal features. The attention fusion mechanism in this paper can automatically learn to give different weights to each feature value of the shallow and deep features. The features significantly impacting the classification result will be given a larger weight. In comparison, the features that do not significantly affect the final classification will be given a smaller weight to improve the detection accuracy and generalization ability.

4.3.3 DNN classifier

We use the attention mechanism to fuse shallow and deep features. In the feature fusion part (as shown in Fig. 3), firstly, we concatenate the shallow and deep features of phase and frequency to obtain the input of length L . Then, to get the weight of each feature, we will input the fully connected layer through the two activation functions for ReLU and Sigmoid. We use the ReLU activation function to enhance the nonlinearity and obtain the weight through Sigmoid. Finally, the input features are multiplied by the weights. Through automatic learning, we give different weights to each value of shallow and deep features. The features that have a significant impact on classification are given greater weights. In comparison, those that are ineffective in classification are given smaller weights to improve detection accuracy.

Table 1 Datasets used in the following experiments

Dataset	Number of original audio	Number of edited audio	Data source
Classical	250	250	Carioca 1 dataset
			Carioca 2 dataset
			New Spanish dataset
GAUDI-DI	251	502	GAUDI dataset

Table 2 Training set, validation set, and testing set on the same dataset and across-dataset experiments

Dataset	Training set	Validation set	Testing set
Same dataset testing on Classical	319	80	101
Cross-dataset testing on GAUDI-DI	400 (Classical)	100 (Classical)	753 (GAUDI-DI)

5 Experimental results and discussion

In this section, we will introduce the dataset, experimental setup, and experimental results. In order to verify the validity of our work, we designed five groups of experiments to verify our contribution: (1) comparison of the method proposed in this paper with the state-of-the-art methods, (2) validation of the fitting coefficient feature, (3) validation of the feature matrix, (4) validation of the deep feature, and (5) validation of the attention mechanism.

5.1 Dataset and experimental setup

In order to verify the effect of the proposed model on different datasets and prove the generalization ability of the proposed model, we use two different datasets as experimental data. They are the classic dataset composed of three public datasets and the GAUDI-DI dataset we collected. The detailed information is shown in Table 1.

In Table 1, the classic dataset we used contains 500 audios and is a mixture of three public datasets, including Carioca, 1, 2, and New Spanish dataset (from two public Spanish language datasets AHUMADA and GAUDI). In order to verify the generalization ability of the model, we established a GAUDI-DI dataset containing 753 audios, selected 251 original audios of about 20 s from the GAUDI dataset, and performed deletion and insertion tampering operations.

The experimental data are divided as shown in Table 2. When testing on the classical dataset, we divide the classical dataset into a training set, a validation set, and a test set with 319 audios in the training set, 80 audios in the validation set, and 101 audios in the test set. When testing with the GAUDI-DI dataset, we will use the classical dataset for training and the GAUDI-DI dataset for testing. The training and validation sets are from the classical dataset, with 400 audios in the training set, 100 audios in the validation set, and 753 audios in the test set.

All the experiments in this paper are based on the TensorFlow 2.0 deep learning framework and performed on NVIDIA GeForce GTX 1080Ti. The specific parameters used in the experiment are as follows: the loss function is binary cross-entropy, and the

optimizer uses Adam, epochs are 100, the batch size is 32, learning rate decay: initial learning rate is 0.001, Halve every 10 epochs.

5.2 Comparison of the proposed method with the state-of-the-art methods

In this experiment, we compared the proposed method in this paper with four baseline methods to verify the effectiveness of the proposed method. The comparison experiments are performed on the same dataset, where F_0 features and SVM classifier are applied in [24], F_1 features and SVM classifier are used in [23], F_f features and SVM classifier are employed in [22], and F_0 , F_1 , and F_f features are fused, and SVM classifier is also utilized in [16]. While F_0 features and F_1 features are related to phase features, F_f features are related to frequency features, and their extraction details are in Sects. 4.1 and 4.2.

As shown in Table 3, the accuracy and F1-score of the proposed method on Classical dataset and GAUDI-DI dataset are higher than the four baseline methods. The best performance among the traditional methods is the one using feature fusion in [16]. Further, the method in this paper improved accuracy by 2–3.3% and F1-score by 1.5–9% on both datasets compared to [16]. This shows that the proposed method obtains the advantage of fused features along with better feature characterization.

It can be seen that all methods perform better on Classical dataset test than on GAUDI-DI dataset. The main reason for this is that the test on GAUDI-DI dataset uses cross-dataset detection, which means the training model is trained with data from Classical dataset and tested with data from GAUDI-DI dataset. The main purpose is to perform generalization performance tests. The experimental results show that although the performance of the proposed method is degraded on the cross-dataset test, it still obtains a good performance, which indicates that the proposed method in this paper has a good generalization performance.

Meanwhile, it can be seen from Table 3 that the accuracy of the frequency feature is significantly lower than that of the phase feature in both experiments. This is because the length of the instantaneous frequency sequence obtained by Hilbert transform is about 18 times that of the phase sequence obtained by DFT transform. (The frequency length of 35 s audio is 37,281, and the phase length is 2055.) Tampering detection only by the average value of ENF sequence variation has excessive information loss. The proposed method has less information loss of deep features obtained through neural networks, and the fusion of ENF phase and frequency, shallow and deep features improves the detection accuracy and generalization ability.

Table 3 Comparison with the state-of-the-art methods

Methods	Classical dataset		GAUDI-DI dataset	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Nicolalde et al. [24]	92.08	92.54	83.53	78.84
Nicolalde et al. [23]	94.06	95.44	83.67	78.90
Esquef et al. [22]	83.17	82.11	79.02	74.18
Wang et al. [16]	95.05	95.41	84.99	79.14
Our method	97.03	96.91	88.31	88.17

5.3 Validation of fitting coefficient features

In this section, we conduct experiments to verify the validity of the fitted coefficient feature, which is a key part of performing deep feature representation learning. The fitting coefficients contain global information of the phase and instantaneous frequency series. The reshape operation of the feature matrix in the deep feature leads to the loss of the general information, so the fitting coefficients are used to compensate for the global information of the deep feature. We fit the ENF phase and frequency series by the Sum of Sines function in MATLAB fitting toolbox, and the number of Sum of Sines function is verified in this section, and the classifiers are SVM, random forest, and XGBoost. The results are shown in Table 4.

Table 4 shows the results of the experiments on both datasets with the fitted coefficients extracted by fitting with 3–8 Sum of Sines functions. When the detection accuracy was verified on Classical, the highest detection accuracy of the instantaneous frequency f' was 92.08% (8 Sum of Sines functions, SVM method). The highest detection accuracy of the phase ϕ was 91.09% (3 Sum of Sines functions, XGBoost method). The highest detection accuracy of transient frequency F was 86.85% (5 Sum of Sines functions, Random Forest (RF) method). The highest detection accuracy of phase ϕ was 82.74% (6 Sum of Sines functions, Random Forest method) when the generalization ability was verified on GAUDI-DI. Since our fitted coefficient features are used as global information compensation for deeper features, the purpose is to obtain higher model generalization ability. The accuracy of 86.59% was also obtained with 6 Sum of Sines functions when the generalization ability was verified GAUDI. Therefore, the fitted narrative of the 6 Sum of Sines function selected in this paper is used to compensate for the global information of the deep features.

The fitted coefficient feature has low dimensionality, less computation, and better detection accuracy. Furthermore, it can reduce the ENF phase and instantaneous frequency sequences of different lengths to the same dimension, which is convenient for automated detection. Therefore, the ENF phase and instantaneous frequency sequences can be downsampled by using the fitting coefficient feature, and the global information of the ENF phase and instantaneous frequency can be obtained with less computation.

Table 4 Detection accuracy of fitting coefficient features (%)

Feature	Classifier	Dataset	3 Sines	4 Sines	5 Sines	6 Sines	7 Sines	8 Sines
Proposed feature f'	SVM	Classical	79.21	79.21	82.18	88.12	88.12	91.09
		GAUDI-DI	77.29	79.28	86.19	86.59	86.32	83.40
	RF	Classical	87.13	85.15	83.17	86.14	87.13	92.08
		GAUDI-DI	83.27	82.74	86.85	86.59	85.13	86.59
	XGBoost	Classical	88.12	88.12	84.16	89.11	87.13	90.10
		GAUDI-DI	76.49	80.08	81.67	80.48	86.32	85.92
Proposed feature ϕ	SVM	Classical	86.14	78.22	86.14	80.20	81.19	79.21
		GAUDI-DI	77.95	75.83	73.04	74.63	73.71	74.50
	RF	Classical	90.10	89.11	90.10	90.10	89.11	90.10
		GAUDI-DI	79.42	79.68	80.08	82.74	82.60	81.14
	XGBoost	Classical	91.09	89.11	90.10	87.13	86.14	88.12
		GAUDI-DI	80.61	77.29	79.55	80.48	81.01	81.41

Bold means the best performance of tampering detection in the same experimental setting

Table 5 Detection performance of feature matrix $F_{m \times m}, P_{n \times n}$

Feature matrix	Classical dataset		GAUDI-DI dataset	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
$F_{m \times m}$	93.07	93.73	79.55	78.33
$P_{n \times n}$	91.09	91.28	77.69	76.15

Bold means the best performance of tampering detection in the same experimental setting

Table 6 Detection performance of deep feature

Deep feature	Classical dataset		GAUDI-DI dataset	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Frequency deep features	94.06	94.32	84.46	81.12
Phase deep features	86.14	85.21	78.88	75.28
Deep features fusion	95.05	95.36	86.96	83.19

Bold means the best performance of tampering detection in the same experimental setting

5.4 Validation of feature matrix $F_{m \times m}, P_{n \times n}$

In this section, the feature matrix $F_{m \times m}, P_{n \times n}$ obtained in Sect. 4.2.2 is validated. The model we used is the DNN classifier added after the last convolution block 3. The experimental results are shown in Table 5.

Table 5 shows the experimental results of the feature matrix $F_{m \times m}$ on classical and GAUDI-DI. The table shows that the frequency feature matrix achieves an accuracy of 93.07% on classical, which is significantly higher than the 83.17% obtained for the frequency shallow feature F_f and 92.08% for the frequency fitting coefficient. The phase feature matrix $P_{n \times n}$ also has a detection of 91.09% and 77.69%. The frequency feature matrix is much larger than the phase feature matrix, and more information about the difference between real audio and tampered audio is obtained from the ENF frequencies through the convolutional neural network training. The detection accuracy can be further improved by fully utilizing the ENF information through the neural network.

5.5 Validation of deep features

This part will verify the validity of frequency and phase depth features, as shown in Fig. 3 (feature 2, feature 3). After the deep features, we perform classification by DNN classifier (two fully connected layers and one dropout layer with 1024, 256 neurons, Dropout rate = 0.2, and finally one Softmax layer). Meanwhile, we conducted deep phase and frequency feature fusion experiments to splice the deep features and then classify them with DNN. The experimental results are shown in following Table 6.

Table 6 shows the classification effect of deep features. It can be seen that the detection effect of frequency deep features on the two datasets is 94.06% and 84.46%, respectively, which is significantly better than that of frequency shallow features. The detection effect of deep phase features is comparable to that of shallow features. In addition, the accuracy of deep phase feature fusion is higher than that of single features, further demonstrating the role of feature fusion in audio tampering detection. Compared with the

shallow feature F_{0lf} , the deep feature fusion has higher accuracy on the GAUDI-DI dataset, indicating that the deep feature has higher generalization ability.

Also, we found that the shallow phase features outperformed the frequency features for classification, while the deep features outperformed the frequency features for classification (as shown in Fig. 4). When the detection accuracy is verified on Classical and the generalization ability is verified on GAUDI-DI, the phase F_1 of the shallow features outperforms the frequency feature F_f . The reason for the different behaviours of the two curves in Fig. 4 is that the data source of Classical dataset is relatively single, while the GAUDI-DI dataset has a more complex data source, resulting in a higher accuracy on Classical dataset.

In contrast, in the deep features, the frequency feature matrix $F_{m \times m}$ outperforms the phase feature matrix $P_{n \times n}$, and the deep frequency features outperform the deep phase features. It can be judged that shallow features and deep features contain different information, and they are complementary. We use the neural network to obtain more details from the ENF, while the shallow features reflect the discontinuity information of the ENF. Therefore, we can further use the complementary characteristics of shallow and deep features to improve the model's classification accuracy and generalization ability.

5.6 Validation of the fusion of shallow and deep features

In this part, the shallow and deep feature fusion methods proposed in this paper (as shown in Fig. 3) are tested. The experimental variables are the final dropout rate and the use of the attention mechanism for feature fusion. The experimental results are shown in following Table 7.

Table 7 shows the experimental results of this paper with feature fusion as the experimental variable. The table shows that the experimental results on two different datasets show that the proposed method achieves the highest detection accuracy and model generalization ability at dropout rate = 0.2. The feature fusion with attention is better than the mechanism without attention. The model generalization ability of the proposed method is significantly better than that of the method in [16]. The fusion of shallow and deep level features by the attention mechanism allows the full exploitation of the ENF phase and instantaneous frequency features. The complementary nature of the features at different levels is exploited to more accurately characterize the changes in

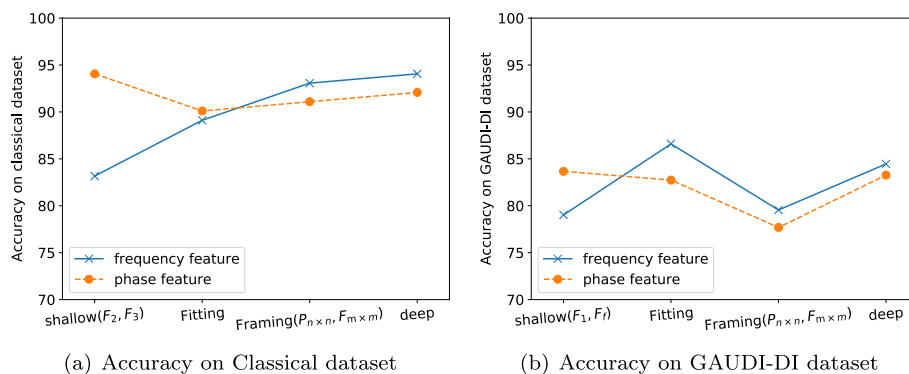


Fig. 4 Accuracy of phase and frequency features, shallow features (F_1, F_f), fitting coefficient (P_{coe}, F_{coe}), input of the convolutional ($P_{n \times n}, F_{m \times m}$), deep phase and frequency features

Table 7 Detection performance of fusion method

Fusion method	Classical dataset		GAUDI-DI dataset		Dropout rate
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	
With attention	95.05	95.41	87.92	87.71	0
	96.04	96.12	87.78	87.46	0.1
	97.03	96.91	88.31	88.17	0.2
Without attention	96.04	96.12	87.65	87.32	0.3
	94.06	94.33	87.38	87.13	0
	95.05	95.41	88.05	87.82	0.1
	95.05	95.41	87.92	87.75	0.2
	95.05	95.41	87.25	87.06	0.3

Bold means the best performance of tampering detection in the same experimental setting

inconsistency produced by tampering operations to natural digital audio. The attention mechanism in this paper can automatically learn can give different weights to each feature value of shallow and deep features. The features significantly impacting the classification result will be given a larger weight. In comparison, the features that have an insignificant effect on the final classification will be given a smaller weight to improve the detection accuracy and generalization ability.

5.7 Discussion

In this section, we conducted five sets of experiments. In experiment 1 (comparison of the proposed method with the state-of-the-art methods), we concluded that the proposed method is better than the state-of-the-art method in [16], and the model generalization ability is significantly better than the baseline methods; in experiment 2 (validation of the fitting coefficient features), we concluded that for the duration of the audio to be measured is 9–35 s, the global information compensation as the deep features; in experiment 3 (validation of feature matrix $F_{m \times m}$, $P_{n \times n}$) and experiment 4 (validation of deep features), we verify the validity of the deep features proposed in this paper; in experiment 5 (validation of the fusion of shallow and deep features), we verify the effectiveness of feature selection by attention mechanism in this paper.

The results show that: (1) Through the fusion of ENF phase and frequency features, audio tampering detection can achieve higher detection accuracy by using different information contained in the ENF phase and frequency in audio. (2) The shallow features contain ENF discontinuity information, while the deep features obtained by the deep learning method contain more ENF details. The complementary feature of the shallow features and the deep features can make the tampering detection method have higher accuracy and generalization ability. (3) The attention mechanism was used for feature fusion, and different weights were assigned to each feature value to suppress invalid features, which further improved the model's performance.

6 Conclusion

This paper proposes an audio tampering detection method based on the fusion of shallow and deep features. Firstly, the phase and instantaneous frequency characteristics of ENF in audio were obtained by DFT and Hilbert transform. Then, we obtained

the shallow layer characteristics reflecting ENF discontinuity through calculation and obtained the deep phase and frequency characteristics through the neural network. Finally, the attention mechanism is used for feature fusion. After dimensionality reduction, the Softmax classifier is used for classification to detect the edited audio. By fusing shallow and deep features, the complementarity of features at different levels is exploited to more accurately describe the changes in inconsistency produced by tampering operations to raw digital audio. Further, attention is used to fuse phase features and frequency features to obtain rich information from audio ENF for tampering detection classification tasks through automatic learning to improve detection accuracy and model generalization. Experimental results show that the proposed method has higher recognition accuracy and generalization ability. Future work will focus on more robust audio tampering detection methods. In addition, detection methods will be designed to locate the location of the audio tamper.

Abbreviations

ENF	Electronic network frequency
STFT	Short-time Fourier transform
MFCCs	Mel-frequency cepstral coefficients
DFT	Discrete Fourier transform
TPSW	Two-pass Split Window
CNN	Convolutional neural network
SVM	Support vector machine

Acknowledgements

The authors acknowledge the comments by anonymous reviewers that helped to improve a preliminary version of the paper.

Author contributions

Equal contribution from all authors. All authors read and approved the final manuscript.

Funding

The research work of this paper was supported by the National Natural Science Foundation of China (Nos. 62177022, 61901165, 61501199), Collaborative Innovation Center for Informatization and Balanced Development of K-12 Education by MOE and Hubei Province (No. xtzd2021-005), and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU2020T010).

Availability of data and materials

Please contact authors for data requests.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 30 September 2021 Accepted: 28 July 2022

Published online: 13 August 2022

References

1. M.A. Qamhan, H. Altaheri, A.H. Meftah, G. Muhammad, Y.A. Alotaibi, Digital audio forensics: Microphone and environment classification using deep learning. *IEEE Access* **9**, 62719–62733 (2021). <https://doi.org/10.1109/access.2021.3073786>
2. C. Zeng, D. Zhu, Z. Wang, Z. Wang, N. Zhao, L. He, An end-to-end deep source recording device identification system for web media forensics. *Int. J. Web Inf. Syst.* **16**(4), 413–425 (2020). <https://doi.org/10.1108/jjwis-06-2020-0038>
3. G. Hua, H. Liao, Q. Wang, H. Zhang, D. Ye, Detection of electric network frequency in audio recordings—from theory to practical detectors. *IEEE Trans. Inf. Forensics Secur.* **16**, 236–248 (2021). <https://doi.org/10.1109/tifs.2020.3009579>
4. C. Zeng, W.Z. Zhu D, Spatial and temporal learning representation for end-to-end recording device identification. *EURASIP J. Adv. Signal Process.* **41**, 1–19 (2021). <https://doi.org/10.1186/s13634-021-00763-1>
5. H. Malik, Acoustic environment identification and its applications to audio forensics. *IEEE Trans. Inf. Forensics Secur.* **8**(11), 1827–1837 (2013). <https://doi.org/10.1109/tifs.2013.2280888>

6. H. Zhao, H. Malik, Audio recording location identification using acoustic environment signature. *IEEE Trans. Inf. Forensics Secur.* **8**(11), 1746–1759 (2013). <https://doi.org/10.1109/tifs.2013.2278843>
7. C. Zeng, D. Zhu, Z. Wang, Y. Yang, Deep and shallow feature fusion and recognition of recording devices based on attention mechanism, in *Advances in Intelligent Networking and Collaborative Systems* (Springer, Cham, 2020), pp. 372–381
8. L. Cuccovillo, S. Mann, M. Tagliasacchi, P. Aichroth, Audio tampering detection via microphone classification, in *15th International Workshop on Multimedia Signal Processing* (2013), pp. 177–182
9. X. Meng, C. Li, L. Tian, Detecting audio splicing forgery algorithm based on local noise level estimation, in *5th International Conference on Systems and Informatics* (2018), pp. 861–865
10. M. Zakariah, M.K. Khan, H. Malik, Digital multimedia audio forensics: past, present and future. *Multimed. Tools Appl.* **77**(1), 1009–1040 (2017). <https://doi.org/10.1007/s11042-016-4277-2>
11. Q. Yan, R. Yang, J. Huang, Detection of speech smoothing on very short clips. *IEEE Trans. Inf. Forensics Secur.* **14**(9), 2441–2453 (2019). <https://doi.org/10.1109/tifs.2019.2900935>
12. Q. Yan, R. Yang, J. Huang, Robust copy–move detection of speech recording using similarities of pitch and formant. *IEEE Trans. Inf. Forensics Secur.* **14**(9), 2331–2341 (2019). <https://doi.org/10.1109/tifs.2019.2895965>
13. J. Chen, S. Xiang, H. Huang, W. Liu, Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. *Multimed. Tools Appl.* **75**(4), 2303–2325 (2014). <https://doi.org/10.1007/s11042-014-2406-3>
14. X. Lin, X. Kang, Exposing speech tampering via spectral phase analysis. *Digital Signal Process.* **60**, 63–74 (2017). <https://doi.org/10.1016/j.dsp.2016.07.015>
15. Z. Xie, Z. Wei, X. Liu, Y. Xue, Y. Yeung, Copy–move detection of digital audio based on multi-feature decision. *J. Inf. Secur. Appl.* **43**, 37–46 (2018)
16. Z. Wang, J. Wang, C. Zeng, Q. Min, Y. Tian, M. Zuo, Digital audio tampering detection based on ENF consistency, in *International Conference on Wavelet Analysis and Pattern Recognition* (2018), pp. 209–214. <https://doi.org/10.1109/icwapr.2018.8521378>
17. A. Hajj-Ahmad, C.-W. Wong, S. Gambino, Q. Zhu, M. Yu, M. Wu, Factors affecting ENF capture in audio. *IEEE Trans. Inf. Forensics Secur.* **14**(2), 277–288 (2019). <https://doi.org/10.1109/tifs.2018.2837645>
18. R. Garg, A. Varna, A. Hajj-Ahmad, M. Wu, “seeing” enf: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing. *IEEE Trans. Inf. Forensics Secur.* **8**, 1417–1432 (2013)
19. G. Hua, H. Zhang, ENF signal enhancement in audio recordings. *IEEE Trans. Inf. Forensics Secur.* **15**, 1868–1878 (2020). <https://doi.org/10.1109/TIFS.2019.2952264>
20. G. Hua, H. Liao, H. Zhang, D. Ye, J. Ma, Robust enf estimation based on harmonic enhancement and maximum weight clique. *IEEE Trans. Inf. Forensics Secur.* **16**, 3874–3887 (2021). <https://doi.org/10.1109/TIFS.2021.3099697>
21. G. Hua, Y. Zhang, J. Goh, V.L.L. Thing, Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Trans. Inf. Forensics Secur.* **11**(5), 1003–1016 (2016). <https://doi.org/10.1109/tifs.2016.2516824>
22. P.A. Esquef, J. Apolinario, L. Biscainho, Edit detection in speech recordings via instantaneous electric network frequency variations. *IEEE Trans. Inf. Forensics Secur.* **9**, 2314–2326 (2014)
23. D. Nicolalde, J. Apolinario, L. Biscainho, Audio authenticity: detecting enf discontinuity with high precision phase analysis. *IEEE Trans. Inf. Forensics Secur.* **5**, 534–543 (2010)
24. D.P. Nicolalde, J.A. Apolinario, Evaluating digital audio authenticity with spectral distances and ENF phase change, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Taipei, 2009), pp. 1417–1420. <https://doi.org/10.1109/icassp.2009.4959859>
25. L. Wang, H. Liang, X. Lin, X. Kang, Revealing the processing history of pitch-shifted voice using CNNs, in *IEEE International Workshop on Information Forensics and Security (WIFS)* (IEEE, Hong Kong, 2018), pp. 1–7. <https://doi.org/10.1109/wifs.2018.8630783>
26. X. Lin, J. Liu, X. Kang, Audio recapture detection with convolutional neural networks. *IEEE Trans. Multimed.* **18**, 1–15 (2016)
27. S. Jadhav, R. Patole, P. Rege, Audio splicing detection using convolutional neural network, in *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (2019), pp. 1–5
28. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, vol. 30 (2017), pp. 1–11

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.