**International Journal of Digital Crime and Forensics**
Volume 14 • Issue 1

# Audio Tampering Forensics Based on Representation Learning of ENF Phase Sequence

Chunyan Zeng, Hubei University of Technology, China
Yao Yang, Hubei University of Technology, China
Zhifeng Wang, Central China Normal University, China*
Shuai Kong, Hubei University of Technology, China
Shixiong Feng, Hubei University of Technology, China

## ABSTRACT

This paper proposes an audio tampering detection method based on the ENF phase and BI-LSTM network from the perspective of temporal feature representation learning. First, the ENF phase is obtained by discrete Fourier transform of ENF component in audio. Second, the ENF phase is divided into frames to obtain ENF phase sequence characterization, and each frame is represented as the change information of the ENF phase in a period. Then, the BI-LSTM neural network is used to train and output the state of each time step, and the difference information between real audio and tampered audio is obtained. Finally, these differences were fitted and dimensionally reduced by the fully connected network and classified by the Softmax classifier. Experimental results show that the performance of this method is better than the state-of-the-art approaches.

## KEYWORDS

Audio Forensics, Deep Learning, Electronic Network Frequency (ENF)

## 1 INTRODUCTION

With the rapid development of Internet communication technology, people get a lot of multimedia information on the Internet every day. Digital audio, as an essential information carrier, occupies a large part of the multimedia content shared and transmitted on the Internet. The emergence of many audio editing software makes the editing operation of digital audio very convenient, and the malicious editing and application of audio by some criminals may lead to some serious consequences (Qamhan et al., 2021). Therefore, there is a growing need for effective editing detection methods, especially when audio is used as an evidence in courtroom trials and in political campaigns or commercial applications.

There are two kinds of tamper detection methods for digital audio (Zakariah et al., 2018). One is the active detection method, which requires embedding watermark and signature in audio in advance to realize audio protection and detection. The other is the passive detection method, which does not need to embed additional information in advance and directly uses standard features contained in digital audio to perform tamper detection.

DOI: 10.4018/IJDCF.302894 *Corresponding Author

In recent years, there have been many studies on passive detection of audio tampering. The audio features used by these passive detection methods include audio statistical features such as voice pitch and formant (Chen et al., 2016; Xie et al., 2018; Yan et al., 2019a), Recording Device Information(Zeng et al., 2020; Zeng et al., 2021), Speaker information(Wang et al., 2020; Wang et al., 2021; Zeng et al., 2018), background noise (Pan et al., 2012) and Electronic Network Frequency (ENF) (Grigoras, 2005; Hua et al., 2016; Rodríguez et al., 2010). ENF is the power line transmission frequency (50 or 60HZ), and ENF is embedded in the audio in the form of buzzing when it is recorded (Hajj-Ahmad et al., 2018). According to the random fluctuation of ENF around the nominal frequency (50 or 60HZ) (Cooper, 2009), audio forensics can be conducted, including timestamp verification (Hua, 2018; Hua et al., 2014), content tampering detection (Esquef et al., 2014; Rodríguez et al., 2010), recording position positioning (Yao et al., 2017; Zheng et al., 2017).

In this paper, we propose a digital audio tampering detection method based on the phase timing characterization of ENF which will help in identifying discontinuities in the ENF signal when audio fragments are inserted or deleted from an existing audio recording. First, we obtain the phase information of ENF by band-pass filtering and Discrete Fourier Transform (DFT) of the audio. Second, we frame the ENF phase to obtain the ENF phase timing representation. Then, we input the ENF phase timing representation into the Bi-LSTM neural network and output the state at each step. Several inputs before and after the edited part jointly determine the presence of ENF mutations. Finally, the tampered audio is detected by fitting a fully connected network and using a Softmax layer. The experimental results show that our method outperforms previous related tampering detection methods.

The contributions of this paper are as follows:

1. The existing ENF-based audio tampering detection methods do not consider the ENF timing characteristics. We use the recurrent neural network to obtain the information of ENF changes over time and finally improve the detection accuracy.
2. When analyzing ENF phase changes for tamper detection, it is necessary to determine whether the ENF phase is abnormal at a certain time. We use the Bi-LSTM network to obtain the forward and reverse information of the ENF phase timing representation to jointly decide whether there is an anomaly, reduce false positives, and obtain higher detection accuracy.
3. Information loss in visual analysis and traditional machine learning methods may cause misjudgment. This paper uses the deep learning method to obtain ENF mutation information from the ENF phase by automatic learning method to reduce information loss and improve detection accuracy.

This paper is divided into five sections, including the introduction. The second section is related work. The third section is the audio tampering detection method proposed in this paper, including: 1. Framework, 2. Feature extraction, 3. Feature processing, 4. Network structure. In the fourth section, we use two data sets and their mixed data sets for experiment and analysis. The last section is the conclusion of this article and the outlook for future research.

## 2 RELATED WORK

The passive detection of the digital audio tamper is to detect whether the audio content is tampered with by using the information contained in the audio under test without a pre-embedded watermark and signature. The audio $s\left(n\right)$ to be detected can be expressed as the following superposition

$$s\left(n\right) = v\left(n\right) + x\left(n\right) + e\left(n\right) + g\left(n\right) \tag{1}$$

where $v(n)$ represents the voice signal itself, $x(n)$ is the ENF component, $e(n)$ is the noise of the equipment, and $g(n)$ is the background noise part. Most of the existing audio passive detection methods of the target are performed using these four features. These include 1. noise of the equipment, 2. background noise, 3. audio speech content, and 4. ENF.

### 2.1 Tamper Detection Research Based on Equipment Noise

Audio recording will leave some traces of recording devices in the audio. The device information in the audio that has not been tampered with should be consistent. Audio forensics can be obtained by comparing reverb in audio or the consistency of information between environment and equipment. In (Capoferri et al., 2020), Capoferri judges whether the reverb in the audio to be tested is consistent in achieving tamper detection according to the feature of different reverb in audio recorded in different environments. Qamhan (Qamhan et al., 2021) classified the microphone and environment of recording, which can be used to analyze the consistency of microphone and environment in audio for tamper detection.

### 2.2 Research on Tamper Detection Based on Background Noise

When audio is recorded in a complex environment, the recorded audio will contain background noise information in the current environment. Tamper operations such as deletion and audio insertion will lead to the discontinuity of background noise in the audio. In order to detect the splicing operation of audio by using background noise and other information in audio, Meng (Meng et al., 2018) detects the heterogeneous splicing tampering of audio by comparing the similarity between variances of syllable background noise. In (Lin & Kang, 2017), Xiaodan Lin uses spectral phase reconstruction to counteract the influence of noise and uses the spectral phase difference before and after audio reconstruction and the spectral phase correlation between adjacent voiced segments to judge whether the audio has tampered.

### 2.3 Research on Tamper Detection Based on Audio Speech Content

The speech in tampered audio will have some inter-frame correlation weakening and mutation of audio features. Chen (Chen et al., 2016) detects tampered audio in the time domain through discrete wavelet packet decomposition and singularity analysis of speech signals for audio tampering operations of insertion and deletion. Saleem (Saleem et al., 2021) inputted the Short-Time Fourier transform (STFT) and Modified Discrete Cosine Transform (MDCT) spectra of audio into the convolutional neural network to identify Spoofed Voices. There is also much research on copy-move Forgery of the copy-paste type. Imran (Imran et al., 2017) used a chaotic theory to make the tampering point possibly exist anywhere in the audio and then detected copy-paste tampering by comparing the differences of the speech spectrum in the voiced part. In (Xie et al., 2018), Zhaozhi Xie combined gammatone feature, Mel-Frequency Cepstral (MFCCs) feature, Pitch feature, and DFT coefficients are four features, and C4.5 decision tree is adopted to realize copy-move tampering detection. Compared with the single feature, this method has a higher detection effect. Qi Yan (Yan et al., 2019b) takes pitch and formant sequences of voiced audio segments as features and realizes copy-move tamper detection through similarity comparison with the threshold. This method has high robustness to joint post-tamper processing operations. After most tampering operations are carried out, some post-processing operations often cover up tampering traces. Therefore, when such post-processing operations are detected in the audio, the audio may have been edited. Qi Yan (Yan et al., 2019a) used the Support Vector Machine (SVM) to test smoothing Operations of editing software based on the local variance of differential signals.

### 2.4 Research on Tamper Detection Based on Electronic Network Frequency

Although there are many audio tamper detection methods, the technology using electronic network frequency ENF is widely used in multimedia forensics (Zakariah et al., 2018). Although ENF is ideally a sinusoidal signal that oscillates at a nominal frequency, the actual ENF signal will fluctuate slightly with the change of energy supply and the load of the power grid (Saleem et al., 2021). When a segment is inserted or deleted from a recording, the ENF of that segment also changes. Therefore, when the ENF frequency or phase information obtained from the audio recording is sudden and discontinuous, it can be judged that the audio has been edited. In recent years, there have been many kinds of research on digital audio tamper detection based on ENF: Guang Hua (Hua et al., 2016) realized timestamp verification and tamper detection based on Absolute Error-map (AEM) between ENF signal audio and database. However, the ENF database is difficult to obtain, and more studies use ENF mutations to detect audio. Esquef (Esquef et al., 2014) detected audio tampering by extracting ENF signals and detecting the consistency of ENF phase changes. Rodríguez (Rodríguez et al., 2010) proposed the TPSW (two-pass Split Window) method to estimate the change degree of ENF background because the tampering operation would cause a sudden change of ENF of tampering point calculated instantaneous frequency by Hilbert transform. Tamper detection by analyzing the mutation point of instantaneous frequency. Reis (Reis et al., 2016) proposed measuring the fluctuation of ENF using esprit-based peak estimation features and automatically detecting the interference of ENF using the support vector machine (SVM). In addition to audio tamper detection, ENF signals are also used for recording location and audio recapture detection. Chowdhury (Chowdhury & Sarkar, 2019) used low Outliers and High Outliers segments based on ENF signal in audio to locate audio using the support vector machine. In (Lin et al., 2016), Xiaodan Lin input ENF sound spectrum as features into the convolutional neural network to train and classify real and recaptured audio.

In order to further increase the detection accuracy and robustness of audio tamper methods based on ENF, some scholars study the characteristics of ENF to obtain better features. Karantaidis (Karantaidis & Kotropoulos, 2021) added a customized lag Window to the Blackman-Tukey acoustic spectrum estimation method to reduce the interference of speech content, making the estimated ENF more accurate. Guang Hua (Hua & Zhang, 2019) proposed a robust filtering algorithm (RFA) to enhance ENF signals in audio, making the extracted ENF signals more accurate. In (Hua et al., 2021), Guang Hua uses RFA to enhance each harmonic component of ENF and finally performs a weighted combination of harmonic components to obtain more accurate ENF estimation.

Existing tamper detection methods include audio-visual analysis or machine learning methods, such as tamper detection using the support vector machine (SVM) training model. These methods may result in information loss, thereby losing the temporal characteristics of ENF. This paper proposes a tamper detection method based on the change characteristics of the ENF phase timing sequence. Due to ENF mutation caused by editing operations, the change degree of ENF in the tamper region will be different from that in other standard regions. We propose a phase timing characterization in which each frame contains the change information of ENF within a period. Each frame was input as a time step into Bi-LSTM, a bidirectional LSTM neural network. LSTM network can fully consider the long-term dependence in time series problems, and Bi-LSTM can jointly decide whether there are abrupt changes and discontinuities in the ENF phase through the information before and after time series. Then, a Softmax classifier is used to detect tampered audio after fully connected network fitting.

Our motivation are as follows:

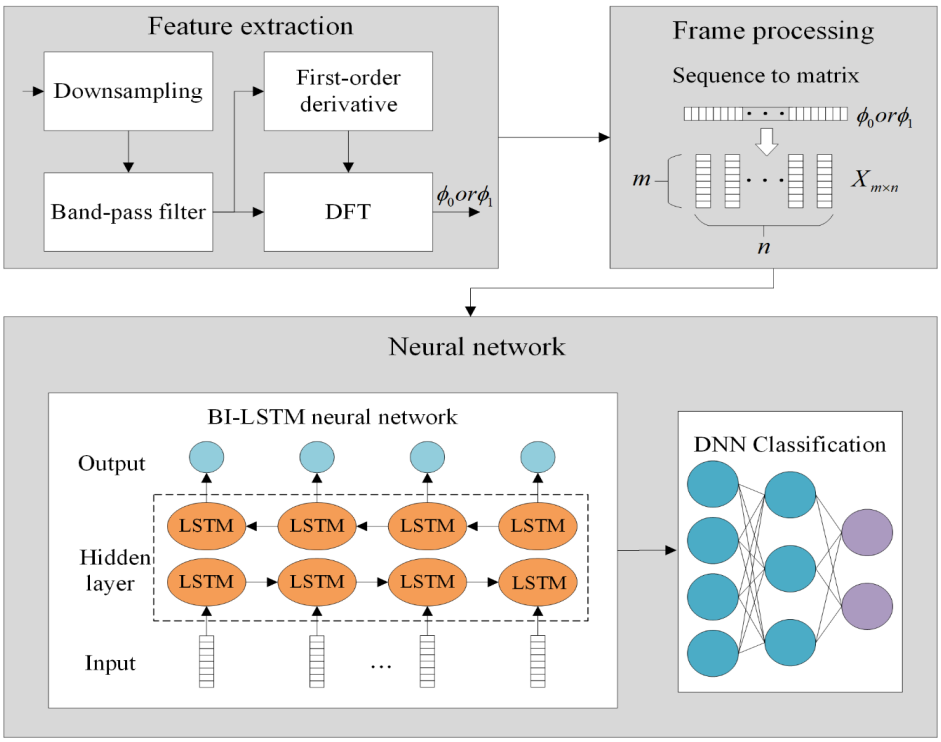
1. The ENF phase change degree information where the audio is edited is different from other areas in the audio. Therefore, when obtaining the state of a region of the ENF phase time series, it is necessary to obtain the information of the previous region. LSTM network can learn the characteristics of long-term dependence of time series to obtain more abundant information about the change of ENF phase over time through the LSTM network.

2. Bi-LSTM network combined with the information before and after the ENF time series is used to determine each time step's state jointly. Through the Bi-LSTM network, ENF phase mutation in audio can be jointly determined by the ENF phase sequence before and after, thus reducing misjudgment.
3. Deep learning method is adopted to automatically learn effective features of audio tamper from a large amount of data, which has a higher degree of automation and detection accuracy than some threshold and visualization methods.

## 3 METHODS

The proposed audio tamper detection method based on the ENF phase sequence characterization is to take the information of ENF phase changes for a period of time as the characteristics of each time step, use the Bi-LSTM network to train and output the state of each time step, and then classify the tamper audio and the original audio through DNN classifier. The method is divided into three parts, as shown in Figure 1:

**Figure 1. The Digital audio tamper detection framework based on ENF phase timing characterization**




1. Feature extraction: ENF components in audio are obtained by down-sampling and band-pass filtering, and then ENF phases $\phi_0$ and $\phi_1$ are obtained by DFT transformation of ENF components and their first derivative.
2. Feature processing: We conducted frame processing on the ENF phase to obtain the ENF phase timing sequence representation $X_{m\times n}$, which has $n$ frames in total, and each frame contains $m$

ENF phase points. Each frame corresponds to the input of each time step of the recurrent neural network to better learn the change information of the ENF phase.

3. Neural network: When the audio is tampered with, the ENF component of the tampered part of the audio will mutate. We use the Bi-LSTM layer to train the ENF phase timing sequence representation $X_{m\times n}$ and output the state of each time step. In this way, the state of each time step can be jointly determined by the time series before and after the time step to obtain the different information containing real audio and tamper audio phases. Then through DNN fitting and dimensionality reduction, finally through Softmax classifier detection tampered audio.

### 3.1 Feature Extraction

The proposed method uses the phase characteristics of ENF, and ENF phases $\phi_0$ and $\phi_1$ can be obtained according to the method in literature (Rodríguez et al., 2010). $DFT^k$ represents the $DFT$ transformation of the signal k derivative, $\phi_0$ represents the phase obtained by $DFT^0$, and $\phi_1$ represents the phase obtained by $DFT^1$. Firstly, the ENF components in audio were obtained by down-sampling and bandpass filtering. Sampling frequencies will be set to 1000Hz and 1200Hz depending on the ENF nominal frequency of 50 or 60Hz. Then bandpass filtering is carried out, and a 10000 order linear zero-phase FIR filter is used for narrowband filtering. The center frequency is ENF standard (50Hz or 60Hz), the bandwidth is 0.6Hz, the passband ripple is 0.5dB, and the stopband attenuation is 100dB. Finally, DFT transformation is performed to obtain ENF phase after obtaining ENF component.

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

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

Where $f\left(*\right)$ represents the approximate derivative operation, and $X_{ENFC}\left[n\right]$ represents the n-th point of the ENF component.

Then, Hanning window $w\left(n\right)$ was used to frame and window $x'_{ENFC}\left[n\right]$. 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 $\frac{1}{60}$ ).

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

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

To obtain the phase $\phi_0$ of ENF and the phase $\phi_1$ of the first derivative of ENF, n-point discrete Fourier transform (DFT) should be executed for each frame signal $x'_N\left[n\right]$ and $X_{ENFC}\left[n\right]$ respectively to obtain $X'\left(k\right)$ and $X\left(k\right)$. Estimated frequency $f_{DFT^1}$ based on the integer index $k_{peak}$ of $\left|X'\left(k\right)\right|$ peak points

$$f_{DFT^1} = \frac{1}{2\pi}\frac{DFT^1\left[k_{peak}\right]}{DFT^0\left[k_{peak}\right]} \tag{4}$$

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

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

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

Now the ENF phase $\phi_0$ of the conventional DTF transformation can be calculated, $\phi_0 = \arg\left[X\left(k_{peak}\right)\right]$. Through Equation (6), $\phi_1$ can be calculated.

$$\begin{cases} \phi_1 = \arctan\left\{\dfrac{\tan\left(\theta\right)\left[1-\cos\left(\omega_0\right)\right]+\sin\left(\omega_0\right)}{1-\cos\left(\omega_0\right)-\tan\left(\theta\right)\sin\left(\omega_0\right)}\right\} \\ \theta \approx \left(k_{DFT^1}-k_{low}\right)\dfrac{\theta_{high}-\theta_{low}}{k_{high}-k_{low}}+\theta_{low} \end{cases} \tag{6}$$

Where, $\omega_0 \approx 2\pi f_{DFT^1}/f_d$, $f_d$ are heavy sampling frequency, $k_{DFT^1} = f_{DFT^1}N_{DFT}/f_d$, $k_{low} = floor\left[k_{DFT^1}\right]$, $k_{high} = ceil\left[k_{DFT^1}\right]$, $floor\left[a\right]$ is the maximum integer less than $a$, and $ceil\left[b\right]$ is the minimum integer greater than $b$. Since the calculated $\phi_1$ has two possible values, $\phi_0$ is used as a reference, and the value closest to $\phi_0$ in $\phi_1$ is selected as the final $\phi_1$.

### 3.2 Feature Processing

After obtaining the ENF phase information in the audio, we conducted frame processing to use the deep learning method better to learn abnormal information such as mutation from the ENF phase with different lengths. The steps are as follows

Step 1: Set the frame length $m$ (phase sample points) and calculate the maximum number of phase points $P_{max}$ in the data.

Step 2: Calculate the number of frames $n$ according to the set frame length $m$ and $P_{max}$, $n = ceil\left(\frac{P_{max}}{m}()\right)$

Step 3: Frame the ENF phase to traverse the phase $\phi_{0,1}$ of all audio data, Calculate the frame shift.

$$overlap = m - floor\left(\frac{length\left(\phi\right)}{n}\right)$$

Step 4: The frame is divided into two parts $X_{m\times n}=\left(X^{frist}_{m\times k},X^{second}_{m\times(n-k)}\right)$, and the frameshift of $X^{frist}_{m\times k}$ frame is one smaller than that of $X^{second}_{m\times(n-k)}$ frame. $k=length\left(\phi\right)-\left(m-overlap\right)\times n$

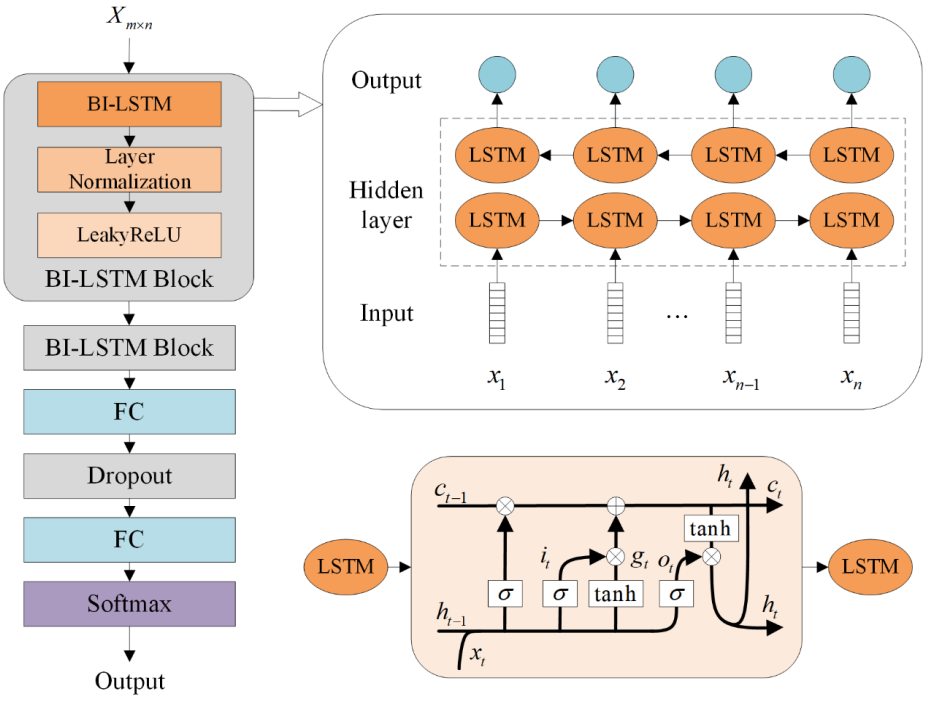
Step 5: Finally, ENF phase sequence characterization $X^{0}_{m\times n},X^{1}_{m\times n}$ can be obtained.

This paper proposes the ENF phase change feature $X_{m\times n}$ to reduce information loss and learn the long-term change information of the ENF phase by using the recurrent neural network. Where $n$ represents a total of $n$ time steps, and each time step contains $m$ phases. These $m$ phase points are the characteristics of the change information of the ENF phase in this period of time.

### 3.3 Neural Network Structure

After filtering, DFT, and frame splitting, we obtained the ENF phase sequence representation $X_{m\times n}$, consisting of $n$ ENF phase frames with dimension $m$. Each frame represents the ENF phase change information within a period of time. As shown in Figure 2, we input it into the Bi-LSTM to obtain the forward and reverse change information of the ENF phase. This paper uses two Bi-LSTM blocks to obtain information about ENF phase changes. The Bi-LSTM block contains a bi-directional LSTM layer, a layernormalization layer, and the leakyrelu activation function. Layer Normalization (LN) (Hou et al., 2019) accelerates network convergence and makes the model more stable. As a result, the LN Layer regularizes and makes the obtained model less likely to be over-fitted. After the Bi-LSTM block, the fitting characteristics of the fully connected layer are used, and the Dropout layer is used to prevent overfitting. Finally, SoftMax is used for classification.

**Figure 2. Bi-LSTM network structure**



### *3.3.1 LSTM and Bi-LSTM Neural Networks*

Bi-LSTM network is a recurrent neural network consisting of the input, hidden, and output layers. As shown in Figure 2, the hidden layer of Bi-LSTM is composed of two LSTM layers. These two LSTM networks process the sequence forward and reverse, respectively, to simultaneously capture the context information of the time series.

The Recurrent neural network (RNN) has the memory function and can save the previous time step information. However, in practical application, RNN often faces gradient disappearance and explosion problems, which results in the limited information that RNN can remember. However, the LSTM network can remember all the information of the sequence when the extended sequence is input, which can solve the long-term dependence problem of time series well. The advantages of LSTM are due to the unique network structure of LSTM. The LSTM network is composed of multiple memory units, as shown in Figure 2. These LSTM cells contain three gate structures: input gate, forgetting gate, and output gate. The combined action of these three gates enables the LSTM network to retain or discard the previous state information and obtain more comprehensive sequence information. At time $t$, the input of the memory unit is the output $h_{t-1}$ and state variable $c_{t-1}$ of the previous time, and the input feature $x_t$ of time $t$. After the following operations are performed by the LSTM cell, the outputs at time $t$ are $h_t$ and $c_t$.

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{7}$$

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{8}$$

$$g_t = \tanh\left(W_{gx}x_t + W_{gh}h_{t-1} + b_g\right) \tag{9}$$

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{10}$$

$$o_t = \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + b_o\right) \tag{11}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{12}$$

In the above formula, $W$ and $b$ represent the values of weights and biases of the neural network, and $\sigma$ and $tanh$ represent sigmoid and Hyperbolic tangent activation functions, respectively. $f_t$ stands for forgetting gate, which determines which information to discard from the cell state. $i_t$ represents the input gate that determines what new information is put into the cell state. $o_t$ indicates the output gate, which determines what the current cell wants to output.

From the structure of the LSTM cell, it can be seen that one-way LSTM only can learn past sequence information. The hidden layer of Bi-LSTM consists of two LSTM layers, which work forward and backward respectively, and obtain the information before and after the time series. In this paper, Bi-LSTM is adopted to make the obtained state at time $t$ contain the information before and after $t$. The ENF information before and after time $t$ can judge whether there are anomalies and mutations jointly.

### *3.3.2 DNN Classifier*

For the information of ENF phase change output by the Bi-LSTM network, we use two fully connected layers to fit it (the number of neurons is 1024,256 respectively, and the activation function is Leaky ReLU). A Dropout layer is added between the two fully connected layers to prevent overfitting (Dropout rate=0.2). Finally, tamper detection is performed through a full-connection layer (the number of neurons is 2, and the activation function is Softmax).

## 4 EXPERIMENT AND ANALYSIS

In this section, the audio tamper detection method based on ENF timing sequence representation proposed in this paper is verified.The contents of this section are as follows:(1) Data set and experimental settings.(2) The method in this paper is compared with the traditional machine learning method. (3) Frame length verification of ENF timing sequence change feature. (4) Verification of multiple temporal feature models. (5) Experimental Analysis.

### 4.1 Data Set and Experimental Settings

To verify the effectiveness of the proposed method, we performed experiments on two data sets (Classical, Gaudi-DI). The Classical dataset mixes three datasets, Carioca1, 2, and New Spanish(from two standard Spanish datasets, AHUMADA and GAUDI). In the GAUDI-DI data set, 251 original audio files were selected from the GAUDI data set, and a total of 753 audio files were obtained through deletion and insertion tampering. The data set details are shown in Table 1

**Table 1. Data set information**

| The data set | Classical | GAUDI-DI | Classical & GAUDI-DI |
|---|---|---|---|
| Edited audio | 250 | 251 | 501 |
| Authentic audio | 250 | 502 | 752 |
| Total audio | 500 | 753 | 1253 |
| Audio time | 9~35s | 16~35s | 9~35s |
| The training set | 319 | 480 | 800 |
| Validation set | 80 | 121 | 201 |
| The test set | 101 | 152 | 252 |

We first ran the experiment on the Classical and Gaudi-DI datasets, then mixed the two datasets and reran the experiment. As shown in Table 1, we divided the data set into the training set, verification set, and test set in the experiment. All experiments in this paper are based on Tensorflow 2.1 deep learning framework, using GPU for NVIDIA GeForce GTX 1080Ti. The experimental parameters

are as follows: Loss function: Binary_Crossentropy; optimizer: Adam; epochs: 400; batch size: 64; Decay: Initial learning rate is 0.001, Halve every 60 epochs.

### 4.2 The Proposed Method is Compared with Traditional Machine Learning Methods

In this section, we will conduct DFT and Hilbert transforms for the filtered ENF components according to the automatic detection method in literature(Wang et al., 2018) respectively to obtain ENF phases $\phi_0$, $\phi_1$, and $f_{hil}$, and then calculate the mean values of the changes of the three features $F_0$, $F_1$, and $F_{hil}$. Finally, an SVM classifier is used for training and testing. At the same time, we set the frame length of the ENF phase sequence characterization frame proposed in this paper as 85 phase points and calculate the total frame number as 25 to obtain features $X^0_{85\times25}$ and $X^1_{85\times25}$. Input $X^0_{85\times25}$ and $X^1_{85\times25}$ respectively into the Bi-LSTM network proposed in this paper for training and testing (see Figure 2). We spliced the features of each time step of feature $X^0_{85\times25}$ and $X^1_{85\times25}$ to obtain feature vector $X^0_{2125}$ and $X^1_{2125}$ with length of 2125, and carried out experiments with Convolutional Neural Network (CNN, 4 convolution layers, 2 maxpooling layers, and finally DNN classifier was used for classification). The experimental results are shown in Table 2 and Figure 3

**Table 2. Classification results of the proposed method and the traditional method**

| Feature | Method | Classical(%) | GAUDI-DI(%) | Classical&GAUDI-DI(%) |
|---|---|---|---|---|
| $F_0$ | SVM (Rodríguez et al., 2010) | 92.08 | 88.16 | 90.48 |
| $F_1$ | SVM (Rodríguez et al., 2010) | 95.05 | 88.16 | 90.48 |
| $F_{hil}$ | SVM (Reis et al., 2016) | 83.17 | 88.16 | 84.13 |
| $F_0F_1F_{hil}$ | SVM (Wang et al., 2018) | 95.05 | 90.13 | 96.03 |
| $X^0_{2125}$ | CNN | 96.04 | 90.13 | 96.03 |
| $X^1_{2125}$ | CNN | 96.04 | 89.47 | 96.43 |
| $X^0_{85\times25}$ | Bi-LSTM | 97.03 | 90.13 | 96.83 |
| $X^1_{85\times25}$ | Bi-LSTM | 97.03 | 90.79 | 97.22 |

**Figure 3. Classification curves of the proposed method and the traditional method**

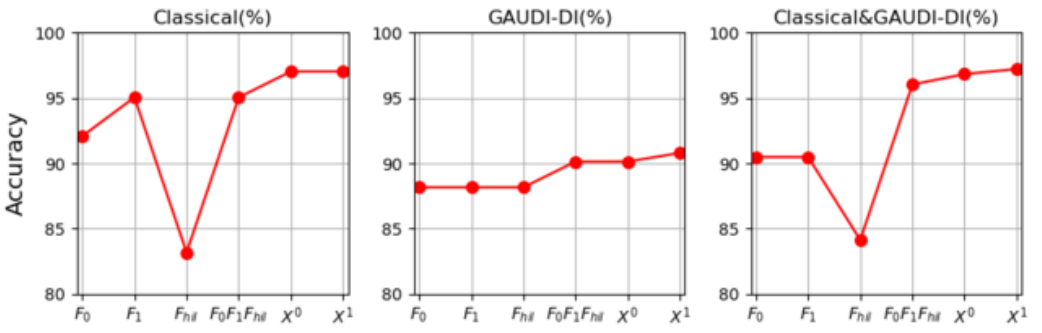



Table 2 and Figure 3 are the classification results of the phase sequence characteristics of ENF proposed in this paper and the automatic detection methods in literature (Wang et al., 2018). As shown from Table 2 and Figure 3, this article proposed that the audio tamper detection method is superior to the traditional machine learning method. In a mixed data sets up to 97.22% accuracy, we use the neural network automatically learning from ENF temporal variation characteristics of better learning to ENF phase mutation information due to tamper with the operation.

Comparing the classification effects of the two methods on three data sets with different data volumes, the Bi-LSTM method proposed by us is superior to the SVM method in the case of the same features. In the case of three data sets, when the feature source is $\phi_0$, the improvement is 4.95%, 1.97%, and 6.35%. When the feature source was $\phi_1$, the elevations were 2.01%, 2.63%, and 6.74%, respectively. Compared with the fusion features in literature(Wang et al., 2018), there are also 1.98%, 0.66% and 1.19% improvements. Compared with the convolutional neural network, the detection accuracy of the Bi-LSTM network used in this paper is superior to CNN in all three data sets. The results show the effectiveness of the detection method proposed in this paper. The ENF timing sequence representation $X_{m\times n}$ information loss is small, and the deep learning method can learn more ENF mutation characteristics to achieve a higher audio tamper detection effect.

### 4.3 ENF Timing Sequence Characterization Verification with Different Frame Lengths

This part will verify the frame length of the ENF phase timing characterization proposed in this paper. According to the maximum number of phase samples obtained from the feature processing part and the most extended audio duration in the data set, a phase sample point of about 0.017s can be calculated. The frame length of ENF phase timing sequence characterization of $\phi_0$ and $\phi_1$ was set to 5-95 samples (the interval was ten samples, 0.085-1.7s, the interval was 0.17s) for experiments. The model used (shown in Fig. 2) consists of two Bi-LSTM blocks followed by two fully connected layers (512,256 neurons). The activation function is Leaky ReLU), and dropout (rate=0.2) is also used between two fully connected layers, which are finally sorted by SoftMax. The experimental results are shown in Table 3 and Figure 4

**Table 3. Accuracy of ENF phase timing sequence characterization for $\phi_0$ and $\phi_1$ phases with different frame lengths (5-95 samples)**

| $m \times n$ of $X_{m\times n}$ | | Classical(%) | | GAUDI-DI(%) | | Classical&GAUDI-DI(%) | |
|---|---|---|---|---|---|---|---|
| $m$ | $n$ | $\phi_0$ | $\phi_1$ | $\phi_0$ | $\phi_1$ | $\phi_0$ | $\phi_1$ |
| 5(0.085s) | 411 | 93.07 | 94.06 | 83.55 | 84.87 | 93.65 | 94.05 |
| 15(0.255s) | 137 | 95.05 | 93.07 | 89.47 | 90.79 | 94.84 | 95.63 |
| 25(0.425s) | 83 | 96.04 | 97.03 | 90.79 | 90.79 | 96.43 | 96.43 |
| 35(0.595s) | 59 | 96.04 | 96.04 | 90.79 | 90.79 | 96.83 | 96.43 |
| 45(0.765s) | 46 | 97.03 | 96.04 | 90.79 | 91.45 | 96.03 | 96.83 |
| 55(0.935s) | 38 | 96.04 | 96.04 | 90.79 | 90.13 | 96.43 | 96.43 |
| 65(1.105s) | 32 | 97.03 | 95.05 | 90.13 | 90.13 | 96.03 | 95.63 |
| 75(1.275s) | 28 | 96.04 | 97.03 | 90.13 | 90.79 | 96.43 | 96.03 |
| 85(1.445s) | 25 | 97.03 | 97.03 | 90.13 | 90.79 | 96.83 | 97.22 |
| 95(1.615s) | 22 | 96.04 | 97.03 | 90.13 | 89.47 | 95.24 | 95.63 |

**Figure 4. Accuracy curves of ENF phase timing sequence characterization on three data sets (5-95 samples).**

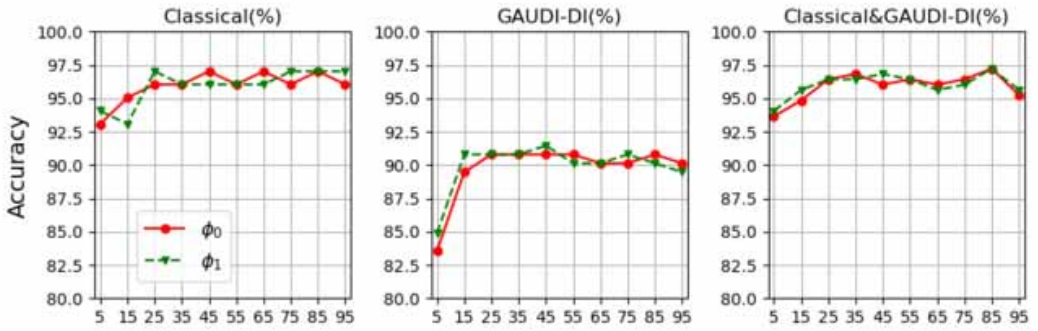



Table 3 shows the experimental results of tamper detection classification on the Bi-LSTM network of ENF phase sequence characterization $X_{m\times n}$ of different frame lengths obtained by using ENF phase $\phi_0$ and $\phi_1$. Figure 4 shows the curve of $X_{m\times n}$ detection accuracy changing with frame length on the three data sets. It can be seen that when the frame length m of $X_{m\times n}$ is 25 phase points (about 0.425s), the detection accuracy is high. When the frame length is 85 phase points (about 1.445s), the accuracy tends to decrease. The proposed method in this paper takes the change information of the ENF phase as the feature of each time step, uses the Bi-LSTM network to train and output the state of each time step, and then classifies tamper audio and original audio by DNN classifier. The results show that the proposed method can better obtain the difference of ENF phase between real audio and tampered audio through the state of each time step, and the number of phase points of each time step should be 25-85 (0.425s~1.445s).

Meanwhile, from Figure 4 of the experiment in the previous section, we can see that the detection accuracy of $F_1$ features obtained by ENF phase $\phi_1$ on the Classical data set is significantly better than that of $F_0$ features obtained by phase $\phi_0$. However, the accuracy of ENF phase $\phi_0$ and $\phi_1$ on the mixture of GAUDI-DI and two data sets is similar. Similarly, it can be seen from the figure in this section that there is little difference in classification effect between ENF phase $\phi_0$ and $\phi_1$ characterization features by using the method proposed in this paper. Phase $\phi_1$ is the high-precision phase feature obtained by the first derivative of the ENF signal. ENF mutation in phase $\phi_1$ has higher precision than phase $\phi_0$, which makes ENF mutation information carried in feature $F_1$ more prominent, making ENF phase $\phi_1$ have a better effect when the amount of data is small (Rodríguez et al., 2010). However, as the amount of data increases, the network has more information, making the difference between real audio and tampered audio more obvious (Najafabadi et al., 2015). However, the ENF phase timing sequence characterization $X_{m\times n}$ information proposed in this paper has less loss. The Bi-LSTM neural network can better learn such differences and ENF phase mutations. Therefore, $\phi_0$ and $\phi_1$ of ENF phases can achieve good results in the three data sets.

### 4.4 Verification of Multiple Recurrent Neural Network Models

This section verifies the timing model used for the audio tampering detection method proposed in this paper. We input the ENF phase timing sequence representations $X_{25\times83}, X_{55\times38}, X_{85\times25}$, with frame lengths of (25(0.425s), 55(0.935s), 85(1.445s) into various recurrent neural networks (RNN, GRU, LSTM, Bi-RNN, Bi-GRU, Bi-LSTM) for training and testing. The model we verified is shown in Figure 2, and only the cyclic neural network layer is replaced when testing different models. The experimental results are shown in Table 4, Table 5, and Figure 5.

**Table 4. ENF time sequence characterization of $\phi_0$ phase with frame length 25,55,85 in different cyclic neural network experiment results**

| | Classical(%) | | | GAUDI-DI(%) | | | Classical&GAUDI-DI(%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Frame length $m$ | 25 | 55 | 85 | 25 | 55 | 85 | 25 | 55 | 85 |
| RNN | 91.09 | 95.05 | 97.03 | 90.79 | 90.79 | 90.13 | 93.65 | 96.43 | 96.43 |
| GRU | 93.04 | 96.04 | 97.03 | 90.79 | 90.13 | 90.79 | 95.63 | 96.03 | 96.03 |
| LSTM | 97.03 | 96.04 | 97.03 | 90.13 | 90.79 | 90.79 | 95.63 | 96.83 | 96.43 |
| Bi-RNN | 91.09 | 95.05 | 96.04 | 90.13 | 90.79 | 90.13 | 94.05 | 96.03 | 96.43 |
| Bi-GRU | 95.05 | 96.04 | 97.03 | 90.79 | 91.45 | 90.13 | 95.24 | 96.03 | 96.43 |
| Bi-LSTM | 96.04 | 96.04 | 97.03 | 90.79 | 90.79 | 90.13 | 96.43 | 96.43 | 96.83 |

**Table 5. ENF time sequence characterization of $\phi_1$ phase with frame length 25,55,85 in different cyclic neural network experiment results**

| | Classical(%) | | | GAUDI-DI(%) | | | Classical&GAUDI-DI(%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Frame length $m$ | 25 | 55 | 85 | 25 | 55 | 85 | 25 | 55 | 85 |
| RNN | 91.09 | 95.05 | 97.03 | 90.13 | 91.45 | 90.79 | 95.63 | 96.03 | 96.43 |
| GRU | 94.06 | 96.04 | 97.03 | 90.13 | 90.79 | 90.13 | 95.63 | 96.03 | 96.43 |
| LSTM | 97.03 | 97.03 | 97.03 | 90.79 | 90.13 | 90.79 | 95.63 | 96.43 | 96.83 |
| Bi-RNN | 93.07 | 96.04 | 96.04 | 90.79 | 90.79 | 90.79 | 95.63 | 95.63 | 96.43 |
| Bi-GRU | 95.05 | 96.04 | 97.03 | 90.13 | 90.13 | 90.79 | 96.03 | 96.03 | 96.83 |
| Bi-LSTM | 97.03 | 96.04 | 97.03 | 90.79 | 90.13 | 90.79 | 96.43 | 96.43 | 97.22 |

**Figure 5. Classification effect curves of $X_{25\times83}$, $X_{55\times38}$ and $X_{85\times25}$ of features of $\phi_0$ and $\phi_1$ in different data sets**

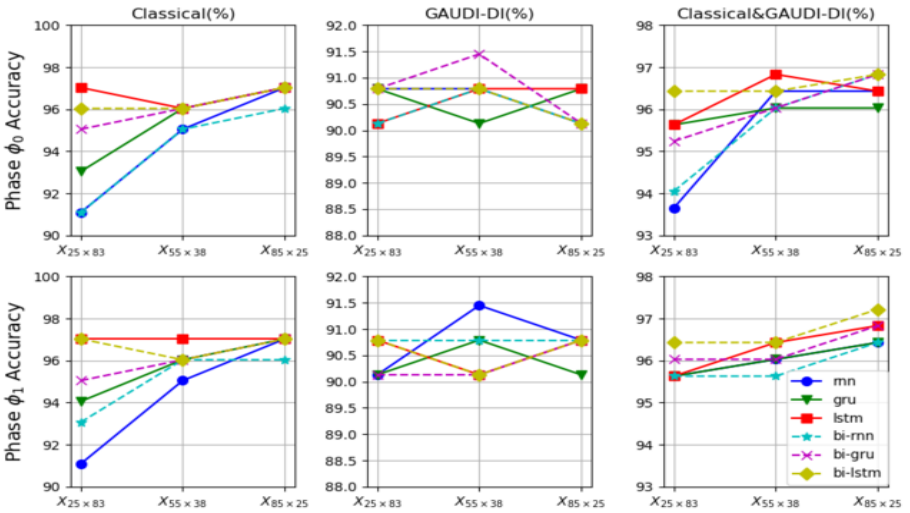



Table 4, Table 5, and Figure 5 show the classification effects of ENF phase timing sequence characterization $X_{m\times n}$ (m=25,55,85) trained with phases $\phi_0$ and $\phi_1$ using different cyclic neural networks. As can be seen from Figure 5, when feature $X_{25\times83}$ is used, the classification effect of LSTM and Bi-LSTM networks is superior to other circulating neural networks. When feature $X_{85\times25}$ is used, the total number of frames is reduced, and the effect of other recurrent neural networks is also significantly improved. ENF phase timing with a frame length of 25 indicates that $X_{25\times83}$ has more frames than $X_{55\times38}$ and $X_{85\times25}$. Since the LSTM network deals with long-term dependence, the phase information of multiple time steps in front of the ENF phase mutation region can make the ENF phase mutation more obvious.

As can be seen from the results of the mixed data set in Figure 5, the classification effect of the three features with different frame lengths in the Bi-LSTM network is more stable than other circulating neural networks. The classification accuracy is better than other neural networks. The bidirectional LSTM network can make the state of each time step jointly determined by the time series before and

after (Xu et al., 2019). The network can more comprehensively obtain the ENF phase difference of authentic audio and tampered audio. It will make the detection more accurate. Therefore, we conclude that the Bi-LSTM network is suitable for training the ENF phase timing sequence characterization proposed in this paper.

### 4.5 Experimental Analysis

We carry out three groups of experiments: 1. Comparison between the proposed method and the traditional machine learning method. 2. Frame length verification of ENF phase timing sequence characterization. 3. Verification of multiple cyclic neural networks. In experiment 4.1, we compared the audio tamper detection method proposed in this paper based on Bi-LSTM with the automatic detection method proposed in the literature (Wang et al., 2018). The results show that feature X proposed in this paper has less information loss. The Bi-LSTM network can learn the ENF phase mutation better, making the proposed method significantly better than the traditional machine learning method. Experiment 4.2 gave ENF phase timing to characterize X with different frame lengths and used LSTM network training for audio tamper detection. The experimental results show that when the audio data is 9~35s, the frame length of the ENF phase timing change feature should be set within 0.425s~1.445s. In experiment 4.3, we used six cyclic neural networks (RNN, GRU, LSTM, Bi-RNN, Bi-GRU, Bi-LSTM) for training and testing. LSTM network is good at dealing with the long-term dependence problem in time series, and the bidirectional cyclic neural network can make the state of each time step jointly determined by the time series before and after. Bi-LSTM network can obtain the information before and after the ENF mutation point and jointly determine whether there is an anomaly (Xu et al., 2019). Therefore, the Bi-LSTM network is more suitable for processing the ENF phase time series characteristics proposed in this paper.

## 5 CONCLUSION

This paper proposes an audio tamper detection method based on the ENF phase and Bi-LSTM. First, the ENF component in audio is obtained by down-sampling and band-pass filtering, and DFT obtains the ENF phase. Then, the ENF phase is divided into frames. Each frame is represented as the ENF phase change degree within a period to obtain the ENF phase sequence characterization. Finally, the Bi-LSTM neural network is used for training. After dimensionality reduction and complete fitting of each time step output by the Bi-LSTM network, the Softmax classifier is used to classify and detect the edited audio. Experimental results show that this method has higher detection accuracy and is better than the existing audio tamper detection methods. Future work will focus on more robust audio tamper detection methods. In addition, audio tamper detection methods will be designed to locate tampered locations in audio.

## ACKNOWLEDGMENT AND FUNDING AGENCY

This research was supported by National Natural Science Foundation of China (No.61901165, 62177022, 61501199), Science and Technology Research Project of Hubei Education Department (No. Q20191406), Hubei Natural Science Foundation (No. 2017CFB683), Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU20ZT010), and Hubei Research Center for Educational Informationization Open Funding (No. HRCEI2020F0102).

## REFERENCES

Capoferri, D. (2020). Speech Audio Splicing Detection and Localization Exploiting Reverberation Cues. *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. doi:10.1109/WIFS49906.2020.9360900

Chen, J., Xiang, S., Huang, H., & Liu, W. (2016). Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. *Multimedia Tools and Applications*, *75*(4), 2303–2325. doi:10.1007/s11042-014-2406-3

Chowdhury, D. (2019). Location forensics analysis using enf sequences extracted from power and audio recordings. https://arxiv.org/abs/1912.09428

Cooper, A. J. (2009). An automated approach to the Electric Network Frequency (ENF) criterion: Theory and practice. *International Journal of Speech Language and the Law*, *16*(2). Advance online publication. doi:10.1558/ijsll.v16i2.193

Esquef, P. A. A., Apolinario, J. A., & Biscainho, L. W. P. (2014). Edit detection in speech recordings via instantaneous electric network frequency variations. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2314–2326. doi:10.1109/TIFS.2014.2363524

Grigoras, C. (2005). Digital audio recording analysis–the electric network frequency criterion. *International Journal of Speech Language and the Law*, *12*(1), 63–76. doi:10.1558/sll.2005.12.1.63

Hajj-Ahmad, A., Wong, C.-W., Gambino, S., Zhu, Q., Yu, M., & Wu, M. (2018). Factors affecting ENF capture in audio. *IEEE Transactions on Information Forensics and Security*, *14*(2), 277–288. doi:10.1109/TIFS.2018.2837645

Hou, L., . . .. (2019). Normalization helps training of quantized LSTM. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc. doi:<ALIGNMENT.qj></ALIGNMENT>10.5555/3454287.3454947

Hua, G. (2018). Error analysis of forensic ENF matching. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. doi:10.1109/WIFS.2018.8630786

Hua, G., Goh, J., & Thing, V. L. L. (2014). A dynamic matching algorithm for audio timestamp identification using the ENF criterion. *IEEE Transactions on Information Forensics and Security*, *9*(7), 1045–1055. doi:10.1109/TIFS.2014.2321228

Hua, G., Liao, H., Zhang, H., Ye, D., & Ma, J. (2021). Robust ENF Estimation Based on Harmonic Enhancement and Maximum Weight Clique. *IEEE Transactions on Information Forensics and Security*, *16*, 3874–3887. doi:10.1109/TIFS.2021.3099697

Hua, G., & Zhang, H. (2019). ENF signal enhancement in audio recordings. *IEEE Transactions on Information Forensics and Security*, *15*, 1868–1878. doi:10.1109/TIFS.2019.2952264

Hua, G., Zhang, Y., Goh, J., & Thing, V. L. L. (2016). Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Transactions on Information Forensics and Security*, *11*(5), 1003–1016. doi:10.1109/TIFS.2016.2516824

Imran, M., Ali, Z., Bakhsh, S. T., & Akram, S. (2017). Blind detection of copy-move forgery in digital audio forensics. *IEEE Access: Practical Innovations, Open Solutions*, *5*, 12843–12855. doi:10.1109/ACCESS.2017.2717842

Karantaidis, G., & Kotropoulos, C. (2021). Blackman–Tukey spectral estimation and electric network frequency matching from power mains and speech recordings. *IET Signal Processing*, *15*(6), 396–409. Advance online publication. doi:10.1049/sil2.12039

Lin, X., & Kang, X. (2017). Exposing speech tampering via spectral phase analysis. *Digital Signal Processing*, *60*, 63–74. doi:10.1016/j.dsp.2016.07.015

Lin, X., Liu, J., & Kang, X. (2016). Audio recapture detection with convolutional neural networks. *IEEE Transactions on Multimedia*, *18*(8), 1480–1487. doi:10.1109/TMM.2016.2571999

Meng, X. (2018). Detecting audio splicing forgery algorithm based on local noise level estimation. *2018 5th international conference on systems and informatics (ICSAI)*. doi:10.1109/ICSAI.2018.8599318

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. doi:10.1186/s40537-014-0007-7

Pan, X. (2012). Detecting splicing in digital audios using local noise level estimation. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/ICASSP.2012.6288260

Qamhan, M. A., Altaheri, H., Meftah, A. H., Muhammad, G., & Alotaibi, Y. A. (2021). Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 62719–62733. doi:10.1109/ACCESS.2021.3073786

Reis, P. M. G. I., Lustosa da Costa, J. P. C., Miranda, R. K., & Del Galdo, G. (2016). ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency. *IEEE Transactions on Information Forensics and Security*, *12*(4), 853–864. doi:10.1109/TIFS.2016.2636095

Rodríguez, D. P. N. (2010). Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security*, *5*(3), 534–543. doi:10.1109/TIFS.2010.2051270

Saleem, S. (2021). Spoofed Voice Detection using Dense Features of STFT and MDCT Spectrograms. *2021 International Conference on Artificial Intelligence (ICAI)*. doi:10.1109/ICAI52203.2021.9445259

Wang, Z. (2020). Robust Speaker Identification of IoT based on Stacked Sparse Denoising Auto-encoders. *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics).* doi:10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00056

Wang, Z. (2021). Robust Speaker Recognition Based on Stacked Auto-encoders. In Advances in Networked-Based Information Systems. doi:10.1007/978-3-030-57811-4_38

Wang, Z.-F. (2018). Digital audio tampering detection based on ENF consistency. *2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. doi:10.1109/ICWAPR.2018.8521378

Xie, Z. (2018). Copy-move detection of digital audio based on multi-feature decision. *Journal of Information Security and Applications, 43*, 37-46. 10.1016/j.jisa.2018.10.003

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment Analysis of Comment Texts Based on BiLSTM. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 51522–51532. doi:10.1109/ACCESS.2019.2909919

Yan, Q., Yang, R., & Huang, J. (2019a). Detection of speech smoothing on very short clips. *IEEE Transactions on Information Forensics and Security*, *14*(9), 2441–2453. doi:10.1109/TIFS.2019.2900935

Yan, Q., Yang, R., & Huang, J. (2019b). Robust Copy–Move Detection of Speech Recording Using Similarities of Pitch and Formant. *IEEE Transactions on Information Forensics and Security*, *14*(9), 2331–2341. doi:10.1109/TIFS.2019.2895965

Yao, W., Zhao, J., Till, M. J., You, S., Liu, Y., Cui, Y., & Liu, Y. (2017). Source location identification of distribution-level electric network frequency signals at multiple geographic scales. *IEEE Access: Practical Innovations, Open Solutions*, *5*, 11166–11175. doi:10.1109/ACCESS.2017.2707060

Zakariah, M., Khan, M. K., & Malik, H. (2018). Digital multimedia audio forensics: Past, present and future. *Multimedia Tools and Applications*, *77*(1), 1009–1040. doi:10.1007/s11042-016-4277-2

Zeng, C. (2018). Stacked Autoencoder Networks Based Speaker Recognition. *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. doi:10.1109/ICMLC.2018.8526953

Zeng, C. (2020). An end-to-end deep source recording device identification system for Web media forensics. *International Journal of Web Information Systems*. 10.1108/IJWIS-06-2020-0038

Zeng, C., Zhu, D., Wang, Z., Wu, M., Xiong, W., & Zhao, N. (2021). Spatial and temporal learning representation for end-to-end recording device identification. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 1–19. doi:10.1186/s13634-021-00763-1

Zheng, L., Zhang, Y., Lee, C. E., & Thing, V. L. L. (2017). Time-of-recording estimation for audio recordings. *Digital Investigation*, *22*, S115–S126. doi:10.1016/j.diin.2017.06.001

*Chunyan Zeng received her PhD degree in Electronic Engineering from South China University of Technology, Guangzhou, P.R. China, in 2013. She is now an associate professor in Hubei University of Technology. Her research interests include compressed sensing and deep learning.*

*Yang Yao was born in September 1997 in China, received the B.Sc. degree in Electrical Engineering and Automation, China, in 2019. He is now a postgraduate student at the Hubei University of Technology. His areas of interest are signal processing, audio forensics.*

*Zhifeng Wang received the BEng degree in Electronic Engineering from China University of Geosciences in 2008, and the PhD degree in Electronic Engineering from South China University of Technology, China, in 2013. He was a joint training PhD student in the computer science department of Carnegie Mellon University, Pittsburg, PA, during 2010 to 2011. He is now an associate professor in the School of Educational Information Technology of Central China Normal University. His research interests include deep learning, data mining, machine learning, signal processing, and digital forensics. He is the corresponding author of this paper and can be contacted via zfwang@ccnu.edu.cn.*

*Shuai Kong was born in China, in May 1998, received the B.C.degree in Electrical engineering and automation, China in 2020. He is now a postgraduate student at Hubei University of Technology, Wuhan, P.R.China. His areas of interests are multimedia security and deep learning.*

*Shixiong Feng was born in China, in Oct 1997, received the B.S. degree in Automation, China in 2019. He is now a postgraduate student at Hubei University of Technology, Wuhan, P.R. China. His areas of interests are multimedia security and deep learning.*

*Yang Yao was born in September 1997 in China, received the B.Sc. degree in Electrical Engineering and Automation, China, in 2019. He is now a postgraduate student at the Hubei University of Technology. His areas of interest are signal processing, audio forensics.*