Electric network frequency based audio forensics using convolutional neural networks

Maoyu Mao, Zhongcheng Xiao, Xiangui Kang, Xiang Li, Liang Xiao

Chapter 14

ELECTRIC NETWORK FREQUENCY BASED AUDIO FORENSICS USING CONVOLUTIONAL NEURAL NETWORKS

Maoyu Mao, Zhongcheng Xiao, Xiangui Kang, Xiang Li and Liang Xiao

Abstract Digital media forensics can exploit the electric network frequency of audio signals to detect tampering. However, current electric network based audio forensic schemes are limited by their inability to obtain concurrent electric network frequency reference datasets from power grids. In addition, most forensic algorithms do not provide high detection precision in adverse signal-to-noise conditions.

This chapter proposes an automated electric network frequency based audio forensic scheme that monitors abrupt mutations of tampered frames and discontinuities in the variations of electric network frequency features. Specifically, the scheme utilizes the multiple signal classification, Hilbert linear prediction and Welch algorithms to extract electric network frequency features from audio signals; the extracted features are passed to a convolutional neural network classifier to detect audio tampering. The negative effects of low signal-to-noise ratios on electric network frequency extraction are addressed by employing extra low-rank filtering that removes voice activity and noise interference. Simulation results demonstrate that the proposed scheme provides better audio tampering detection accuracy compared with a benchmark method, especially under adverse signal-to-noise conditions.

Keywords: Audio forensics, electric network frequency, neural networks

1. Introduction

Audio editing software is often used by malicious actors to reduce the reliability of judicial evidence and defeat intellectual property protection. Audio tampering detection methods mostly rely on fingerprint information embedded in audio signals. Since fragile watermarks cannot assist in detecting private audio signal tampering [1], passive forensic schemes

based on extracted audio features can provide lightweight solutions. Researchers have developed detection methods based on local noise levels of audio signals [18] and voice activity detection [10]. The electric network frequency (ENF) of audio signals demonstrates that power grid features are applicable to digital media forensics [5]. Specifically, electric network frequency signals can be used to verify recording features such as time and location [7, 21], detect synchronization between audio and video data [20] and verify the authenticity of multimedia [14].

Electric network frequency based audio tampering detection techniques can verify if audio recordings have been edited at low computational cost. Ideally, the grid signal is a real sinusoid that fluctuates around its nominal value of 50 Hz or 60 Hz. Given that control mechanisms and power supply parameters are different in different parts of the world, electric network frequency signals display different fluctuations and peak frequency transformations. When signal-to-noise ratio (SNR) conditions are poor, disturbances near the electric network frequency component may be confused with the peak corresponding to the true electric network frequency [12]. Furthermore, due to legal restrictions, it is difficult to obtain concurrent reference datasets of power systems [7]. Additionally, many edit detection schemes based on electric network frequency variations adjust the classification thresholds manually. Although some automated tampering detection schemes do not rely on concurrent power reference datasets, new techniques are required to improve detection accuracy and reduce computational costs.

This chapter proposes an electric network frequency based audio forensic scheme that detects tampering. The scheme assumes a signal model containing the electric network frequency component, where the background noise is low enough to ensure that the electric network frequency signal is the energy-dominant signal around the nominal frequency. Audio tampering is detected without using concurrent reference electric network frequency signals from power networks. The scheme applies two-stage – low-rank and bandpass – filtering to purify electric network frequency signals in a narrow spectral vicinity and compensate for time delays in order to obtain accurate estimates of the real-time edit locations. Based on the sensitivity of electric network frequency features to phase discontinuity changes, variations in the electric network frequency based features extracted from the multiple signal classification, Hilbert linear prediction and Welch algorithms are combined as eigenvectors and input to an automatic classifier. A convolutional neural network (CNN) is employed in the audio tampering detection scheme to improve the generalization ability in practical situations. Simulation results demon-

strate that the proposed audio tampering detection scheme has good accuracy and an expanded application scope.

2. Related Work

Hua et al. [8] have discussed the limitations of electric network frequency based tampering detection systems and the challenges posed by noise and interference. Several electric network frequency extraction algorithms such as the short-time Fourier transform and time recursive iterative adaptive algorithms are incorporated in instantaneous frequency estimation techniques to achieve high-precision extraction by measuring the maximum energy or weighted energy recorded from the average frequencies of spectrograms [6].

A systematic assessment of parametric and non-parametric extraction techniques for electric network frequency signals has demonstrated that time-domain-based extraction algorithms are susceptible to frequency anomalies caused by sudden changes in noise or speech activity [11]. In addition, parametric algorithms such as the multiple signal classification and Welch algorithms can improve resolution frequency estimation of sinusoidal signals by using fewer data series than spectrogram-based extraction algorithms.

An electric network frequency extraction scheme proposed by Lin and Kang [12] applies robust principle component analysis to remove noise interference and purify the electric network frequency when signal-to-noise conditions are poor. It adopts the Hilbert linear prediction algorithm to capture the electric network frequency from fewer audio recordings in an efficient manner.

Nicolalde Rodriguez and Apolinario [16] have developed a digital audio authenticity evaluation scheme that detects electric network frequency phase transitions and leverages the spectral distance using an adaptive filter as a linear indicator. An electric network frequency based edit detection scheme for speech recordings designed by Esquef et al. [3] yields low equal error rate (EER) values by comparing electric network frequency variations around the nominal frequency with the upper limit of the normal variations observed in an unedited signal. Hua et al. [9] have analyzed the absolute error map between an electric network frequency database and test electric network frequency signals to perform timestamp verification and detect tampering via insertion, deletion and splicing attacks with image erosion.

Nicolaide Rodriguez et al. [17] have also developed an automated authenticity detection scheme for audio recordings via phase analysis of high-order electric network frequency harmonics. Reis et al. [19] have

designed an adulteration detection scheme for audio recordings that integrates the kurtosis features of electric network frequency signals in rotational invariance techniques and Hilbert linear prediction in poor signal-to-noise conditions to autonomously classify audio recordings using a support vector machine. Although kurtosis extraction speeds up the classification, some characteristic information is lost.

Wang et al. [22] have developed a detection scheme that applies discrete Fourier transforms of audio signals to achieve instantaneous phase estimation using a support vector machine classifier. However, the accuracy of the scheme is unsatisfactory and the cost of using a support vector machine to evaluate the decision function is linearly related to the number of training samples. This results in high computational costs for large datasets.

Researchers have also applied convolutional neural networks to analyze audio recapture [13] and perform median filtering [2]. However, no research has applied convolutional neural networks to electric network frequency based audio tampering forensics.

3. System Model

Figure 1 shows a schematic diagram of the proposed audio forensic scheme, which uses a convolutional neural network in conjunction with the multiple signal classification (MUSIC), Hilbert linear prediction and Welch algorithms.

The system initially reduces the sampling rate of an audio signal under test $x(m)$. Let ω_0 be the nominal electric network frequency. According to convention, the new sampling frequency f_s is adjusted to 20 times the nominal frequency ω_0. Therefore, the sampled signal $x_{ds}(n)$ where $0 < n \leq m$ is obtained using a 1,000 Hz or 1,200 Hz sampling frequency.

The low-rank structure of the electric network frequency signal in the short-time Fourier transform (STFT) domain is leveraged to separate grid signals from interference by robust principal component analysis (RPCA). Let X_{ds} be the amplitude spectrum of the sampled signal $x_{ds}(n)$. Then, the robust principal component analysis objective is given by:

$$\min_{\hat{X}_C, X_E} \operatorname{rank}\left(\hat{X}_C\right) + \lambda \left\| X_E \right\|_0 \quad \text{s.t.} \quad \hat{X}_C + X_E = X_{ds} \tag{1}$$

where $\| \cdot \|_0$ is the L0-norm, $\lambda > 0$ is a parameter that trades off the low-rank part with the electric network frequency component $\hat{X}_C$, and X_E is the sparsity part containing the impulse noise and speech activity signal.

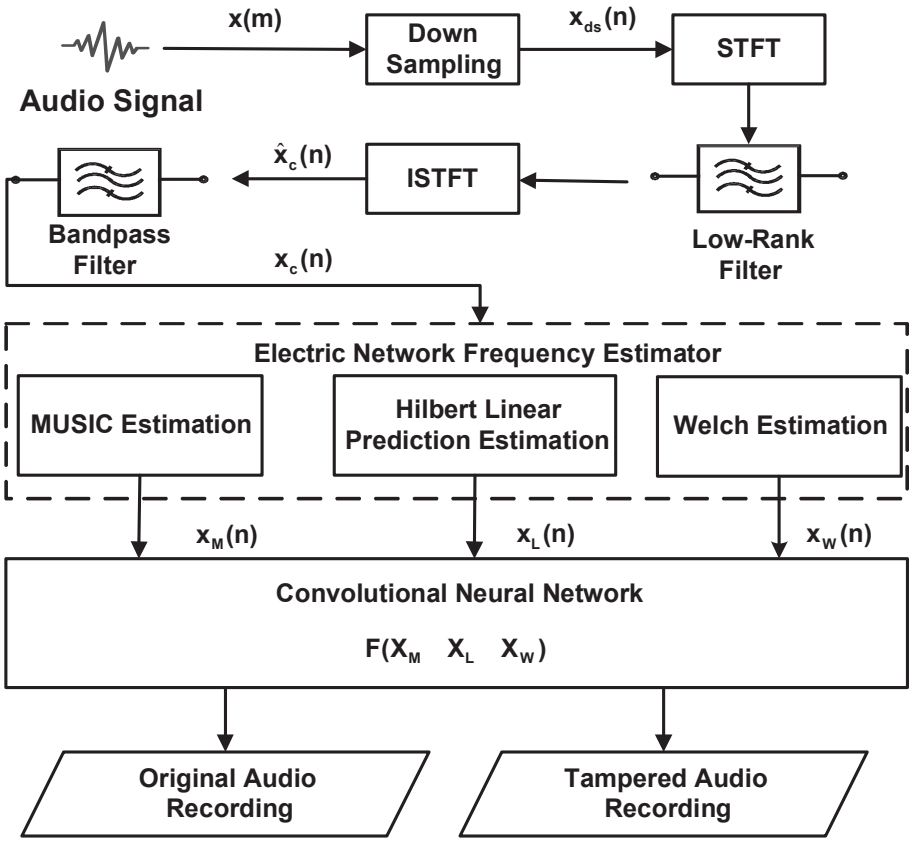

Figure 1. Proposed audio forensic scheme.

Due to the non-convex optimization objective, a relaxation is applied to Equation (1) according to [23]. Thus, the principal component analysis objective becomes:

$$\min_{\hat{X}_C, X_E} \left\| \hat{X}_C \right\|_* + \lambda \left\| X_E \right\|_1 \quad \text{s.t.} \quad \hat{X}_C + X_E = X_{ds} \tag{2}$$

where $\| \cdot \|_*$ is a nuclear norm and $\| \cdot \|_1$ is the L1-norm.

Next, $\hat{X}_C$ is derived by the augmented Lagrange multiplier method [12] and the inverse short-time Fourier transform (ISTFT) is employed to determine the low-rank filtered signal sequence denoted by $\hat{x}_C(n)$.

Grid signals with electric network frequency components $x_C(n)$ are insulated from interference falling into the low-order space by filtering $\hat{x}_C(n)$. Instead of a finite impulse response filter, a fourth-order elliptic filter is adopted with a phase that is approximately linear and adjacent

to the bandpass region. This reduces the computational complexity and computational costs.

The grid signal $x_C^k(n)$ obtained from the audio recording k after two-stage filtering is similar to a narrow-band pseudo-sinusoidal signal. To simplify the presentation, the audio recording index k in the superscript is omitted. Accordingly, the time-domain representation of the electric network frequency signal in the interval is modeled as:

$$x_C(n) = a\cos\left(2\pi\frac{x_F(n)}{f_s}n + \phi\right) \quad 1 \leq n \leq L \tag{3}$$

where L is the length of time-domain signal, $x_F(n)$ is the electrical network frequency to be estimated, f_s is the sampling frequency and a and ϕ are related to the magnitude and phase, respectively.

4. ENF-Based Forensics with CNN

During electrical network frequency extraction, the captured grid signal $x_C(n)$ is typically divided into Y time frames of fixed-length l containing overlapping portions where $Y \in \{L/l\}_{0\leq l\leq L}$. A frequency estimation algorithm is then used to obtain the components of the electrical network frequency characteristics in frame i where $1 \leq i \leq Y$.

The MUSIC algorithm is a subspace spectrum estimation algorithm based on feature structure decomposition. The algorithm decomposes the covariance matrix of a signal sequence into a singular value. By constructing the orthogonal signal and noise subspaces, the algorithm provides spatial spectral functions that can be used to estimate electrical network frequency features. Since electrical network frequency signals contain one real sinusoid, two complex frequency sinusoids are embedded in the white noise P for electrical network frequency signals.

The algorithm first computes the $M \times N$ sample data matrix $\mathbf{A}$ based on the power grid signal $X_C^{(i)}$ of audio frame i:

$$\mathbf{A} = \Big[\mathbf{a}_C(1)\ \mathbf{a}_C(2)\ldots\mathbf{a}_C(N-1)\Big]^T \tag{4}$$

where $\mathbf{a}_C(n) = [x_C(n), x_C(n+1), \ldots, x_C(n+M-1)]^T$, M is the order of the covariance matrix that is chosen to be larger than P and $M \in \left[\frac{N}{3}, \frac{2N}{3}\right]$[6].

Next, the eigenvalue decomposition of the auto-covariance matrix $\mathbf{R} = \frac{1}{N}\mathbf{A}^H\mathbf{A}$ is computed. Since the signal and noise are independent, the covariance can be decomposed and the space comprising the eigenvectors corresponding to the large eigenvalues $(\mathbf{q}_1\ \mathbf{q}_2\ldots\mathbf{q}_P)$ is the

signal subspace $\mathbf{S}$. Also, the space comprising the eigenvectors corresponding to the small eigenvalues $\left(\mathbf{q}_{P+1}\ \mathbf{q}_{P+2}\cdots\mathbf{q}_{P+M}\right)$ is the noise subspace $\mathbf{G}_n$.

The following assumptions are made for the complex form of the observation model described above:

- Different $x_C^{(i)}(n)$ signals are linearly independent of each other.
- The additive noise $u(n)$ is the complex noise with zero mean additive, uncorrelated and the same variance σ_u^2.

Given the orthogonality property of the white noise eigenvectors and signal steering vectors $\mathbf{v}\left(X_C^{(i)}\right)$, the signal steering vectors $\mathbf{v}\left(X_C^{(i)}\right)$ can be written in complex form as:

$$\mathbf{v}\left(X_C^{(i)}\right) = \left[1, e^{j2\pi X_C^{(i)}}, e^{j4\pi X_C^{(i)}}, \ldots, e^{j2(M-1)\pi X_C^{(i)}}\right]^H \tag{5}$$

where $j = \sqrt{-1}$ and $[\,\cdot\,]^H$ denotes the conjugate transposition.

The pseudo-spectral function P_{MU} is computed as:

$$P_{MU} = \frac{1}{\mathbf{v}^*\left(X_C^{(i)}\right)\mathbf{G}_n^i\mathbf{G}_n^{i*}\mathbf{v}\left(X_C^{(i)}\right)} \tag{6}$$

where * is the element conjugate.

Ultimately, the estimated electrical network frequency X_{MU} is obtained by searching for the spectral peak of the spatial-spectral function P_{MU}. The MUSIC algorithm estimates a fixed parameter for each frame, which is the best electrical network frequency value in the least mean square sense for a given signal sequence.

The Hilbert linear prediction extraction algorithm is more sensitive to sharp phase changes than the MUSIC algorithm. However, the MUSIC algorithm is more robust to noise interference. According to Equation (3), the electrical network frequency value $\hat{h}_C(n)$ can be estimated by the transient phase change of the Hilbert transform from the real-valued estimate $x_C(n)$ as follows:

$$\hat{h}_C(n) = x_C(n) + jH\Big\{x_C(n)\Big\} \tag{7}$$

where $j = \sqrt{-1}$ and H is the Hilbert operator.

Since the analytical version of a pseudo-sinusoidal signal is equivalent to the real-valued signal with respect to $x_F(n)$, the linearly predictable property can be applied to the complex model. This yields:

$$\hat{h}_C(n) = ae^{\left(j2\pi\frac{x_F(n)}{f_s}n+\phi\right)} = \beta_1\hat{h}_C(n-1) \tag{8}$$

where $\beta_1 = e^{\left(j2\pi\frac{x_F(n)}{f_s}\right)}$ is the first-order prediction coefficient.

The signal entry $s(n) = x_F(n) + u(n)$ is then obtained by adding the additive complex noise $u(n)$.

Given the assumption that additive complex noise is always equivalent, the approximation $s(n) \approx \beta_1 s(n-1)$ is obtained according to Equation (8). Extending this equation to the entire audio recording yields: $\mathbf{S}_1 \approx \beta_1 \mathbf{S}_2$ where $\mathbf{S}_1 = [s(n-1), s(n-2), \ldots, s(n)]_{0<n\leq m}$ and $\mathbf{S}_2$ is the sequence with one sample shift from $\mathbf{S}_1$.

Therefore, the crux of electrical network frequency estimation is to minimize the weighted linear prediction error in the minimum squared sense as follows:

$$\min J(\beta_1) = \mathbf{e}^T \mathbf{W} \mathbf{e} = (\mathbf{S}_2 - \beta_1 \mathbf{S}_1)^H \mathbf{W} (\mathbf{S}_2 - \beta_1 \mathbf{S}_1) \tag{9}$$

where $\mathbf{W}$ is a symmetric weighting matrix, H is the conjugate transposition operator and $J(\beta_1)$ is the total cost function denoted by the weighted squared error $\mathbf{e}$.

The symmetric weighted matrix $\mathbf{W}$, which is obtained by Markov estimation, is given by:

$$\mathbf{W} = \begin{bmatrix} 1+\|\beta_1\|^2 & -\beta_1 & 0 & 0 & \ldots & 0 \\ -\beta_1^* & 1+\|\beta_1\|^2 & -\beta_1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & -\beta_1^* & 1+\|\beta_1\|^2 & -\beta_1 \\ 0 & 0 & \vdots & 0 & -\beta_1^* & 1+\|\beta_1\|^2 \end{bmatrix}^{-1} \tag{10}$$

where * and $[\cdot]^{-1}$ are the element conjugate and matrix inverse, respectively.

Upon setting the differential in Equation (9) to zero, the prediction coefficient β_1 is given by:

$$\beta_1 = \frac{{\mathbf{S}_1}^H \mathbf{W} \mathbf{S}_2}{{\mathbf{S}_1}^H \mathbf{W} \mathbf{S}_1} \tag{11}$$

Equations (9) through (11) reveal that the computation of β_1 is an iterative process. Having obtained β_1, the Hilbert linear prediction of the electrical network frequency X_L is computed as:

$$X_L = f_s \frac{1}{2\pi} \angle(\beta_1) \tag{12}$$

For consistency with other characteristics, X_L is divided into fragments and the maximum value of the absolute values is taken as the i^{th} segment electrical network frequency estimate $X_L^{(i)}$.

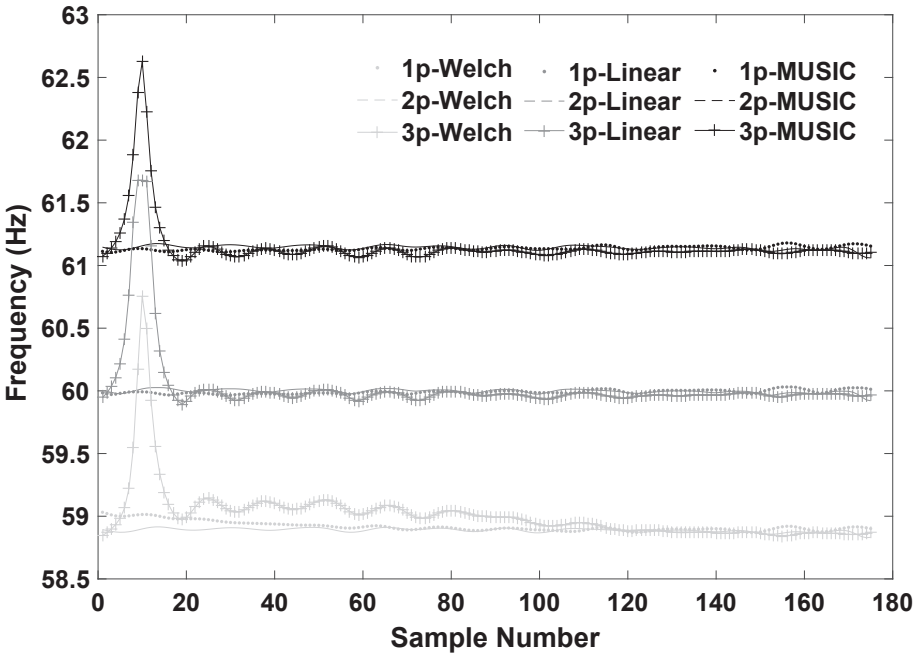

Figure 2. Electrical network frequency fingerprints in three recording fragments.

The Welch algorithm is an improved periodogram method. The algorithm reduces noise in the estimated power spectrum by enhancing the frequency resolution, yielding the largest maximum correlation coefficient around the nominal frequency compared with the MUSIC algorithm and other methods.

The Welch estimate is obtained from the power spectral density. The algorithm divides each audio recording into overlapped segments multiplied by a Hamming window. The frequency sample w, which corresponds to the maximum periodogram value, is extracted as the Welch-based electrical network frequency estimate denoted by $X_W^{(i)}$. Next, a quadratic interpolation is employed to fit the quadratic model of w. The Welch algorithm with a Hamming window improves the spectral distortion caused by the large-side lobe of the rectangular window, yielding an accurate electrical network frequency estimate X_W that is not affected by interference.

Figure 2 clearly shows the electrical network frequency based fingerprints used to verify the effectiveness of the proposed audio forensic scheme. The fingerprints are located in three recording fragments named $1p$, $2p$ and $3p$ at 60 Hz with slight offsetting for easy viewing. The three

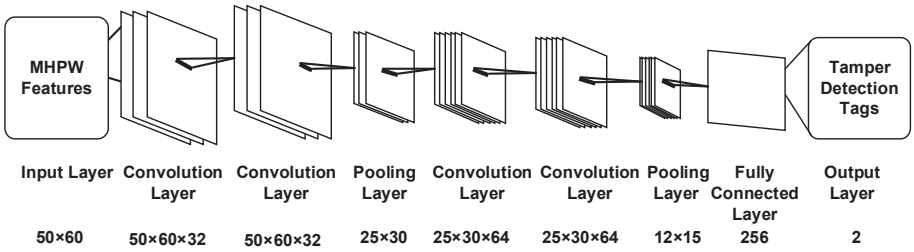

Figure 3. Network architecture of the proposed audio forensic scheme.

recording fragments, all with the same number of samples, are derived from the MUSIC, Hilbert linear prediction and Welch (MHPW) feature estimation algorithms.

The forensic fingerprints simultaneously display high stability and sensitivity to tampering operations. For example, the fingerprints in recording fragments $1p$ and $2p$ show stable pseudo-sinusoidal fluctuations whereas the fingerprints in the tampered fragment $3p$ show sensitive mutations.

Instead of manually determining the threshold, a novel deep learning approach is applied to identify tampered audio recordings. The electrical network frequency signals extracted by the three algorithms are directly modeled as features to avoid information loss when extracting the representative values of features.

Three-dimensional feature vectors $\mathbf{F} = \left[X_L^k, X_{MU}^k, X_W^k\right]_{0<k\leq N}$ are obtained from the N audio recordings. When the lengths of the recordings in the audio dataset are different, the nominal frequency of 50 Hz or 60 Hz is applied to fill the feature vectors to the same length. The three feature channels with the same length constitute the input layer of the neural network structure.

Figure 3 shows the network architecture of the audio tampering detection scheme. The convolutional neural network model has four convolution layers, two pooling layers, one full connection layer and an input layer and output layer. Before processing the features, min-max normalization is used to amplify the differences and variation rules of the features. Next, given the overlaps of the adjacent electrical network frequency components, convolution is used to refine the energy changes in the electrical network frequency signals, which improves the detection accuracy. Finally, the tag distribution obtained by the convolutional neural network model is used to compute the detection performance metrics.

5. Experiments and Results

This section describes the simulation experiments and the results obtained.

5.1 Experimental Setup

The experiments were performed on Matlab and Python 3.6 platforms with the scikit-learn package.

The electrical network frequency based features were extracted from two classical audio databases. The first was the Carioca 1 database [17], a telephone recording database of the public switched telephone network containing 16-bit mono waves at a 44.1 kHz sampling rate and coded by pulse code modulation with an electrical network frequency component around 60 Hz, The second was the Spanish Speech database [4] sampled at 16 kHz with a nominal electrical network frequency component around 50 Hz. The databases each contain 100 original voice audio recordings and 100 edited versions of the original voice audio recordings.

Simulations were performed to evaluate the performance of the audio forensic scheme with $N = 400$ audio recordings. Each recording was divided into time frames of length $l = 1$ second with an overlap of 0.5 seconds. In the simulations, 70% of the original audio recordings and tampered audio recordings were randomly chosen to train the convolutional neural network. The remaining 30% of all the recordings were randomly-chosen for the testing set. Distributing the data into training and testing datasets in this manner ensured that every portion of the data would be more representative. The data randomness had to be high due to the large number of parameters and strong learning ability of the convolutional neural network, and so that the random gradient descent optimization function did not get stuck in a local minimum.

Multiple evaluations were performed to achieve fair comparisons with the benchmark strategy proposed by Reis et al. [19]. The detection error tradeoff (DET) curves were obtained by plotting the false negative rate (FNR) versus false positive rate (FPR) curves for various thresholds [15]. In general, as the false positive rate increases to 100%, the false negative rate decreases, and vice versa. The equal error rate is the point at which the false negative rate and false positive rate are equal. The overall error rate (OER) is computed as the average of the false negative rate and false positive rate.

Table 1. Overall error rates for combinations of fusion features.

Feature	Overall Error Rate
MUSIC	7.5%
Hilbert Linear Prediction	5.1%
Welch	6.3%
MUSIC + Hilbert Linear Prediction	6.3%
Hilbert Linear Prediction + Welch	4.4%
MUSIC + Welch	6.5%
MUSIC + Hilbert Linear Prediction + Welch (MHPW)	3.2%

5.2 Detection Performance

Table 1 shows the overall error rates for feature vectors obtained by combining fusion features. Combining all three features (MUSIC, Hilbert linear prediction and Welch (MHPW)) yields the lowest overall error rate of 3.2% compared with using any one feature or any two features. When all three features are used together, the overall error rates fall by 4.3%, 1.9% and 3.1%, respectively, from the overall error rates when the MUSIC, Hilbert linear prediction and Welch algorithms are used alone.

However, the overall error rates obtained for the mixed features extracted by two of the three algorithms may be suboptimal to those extracted by a single algorithm; this is due to the cancellation of the sharp peak features of the two algorithms. For example, the overall error rate for the fusion features extracted by the MUSIC and Welch algorithms is reduced by 1.0% compared with the overall error rate of the features extracted by the MUSIC algorithm alone. The proper choice of features plays an important role in the accurate detection of audio tampering.

Table 2. Overall error rates for various classifiers.

Classifier	Overall Error Rate
Neural Network	9.1%
Random Forest	9.1%
Decision Tree	7.3%
Logistic Regression	6.7%
Support Vector Machine	4.2%
Convolutional Neural Network	3.2%

Table 2 shows that the proposed scheme using the convolutional neural network with MUSIC, Hilbert linear prediction and Welch features

Table 3. Cross-domain evaluations of the audio databases.

Training Database	Testing Database	Overall Error Rate
Carioca 1	Spanish Speech	4.3%
Spanish Speech	Carioca 1	4.5%

has the lowest overall error rate compared with the other classifiers. For example, the overall error rate is 5.9% less than that obtained by the neural network scheme and is 1.0% less than that obtained by the support vector machine scheme. Additionally, the detection performance of the proposed scheme using the convolutional neural network with the MUSIC, Hilbert linear prediction and Welch features has an overall error rate that is 1.3% less than that obtained by the benchmark strategy with a support vector machine described in [19].

Table 3 shows the results of cross-domain evaluations when the Carioca 1 and Spanish Speech databases were used for training and testing, respectively, and vice versa. Using the combination of MUSIC, Hilbert linear prediction and Welch features trained with the Carioca 1 database yields slightly better prediction results (4.3%) compared with when the Spanish Speech database was used for training (4.5%). This could be because the extracted features of Caricoa 1 are more obvious, which renders the trained model more representative and the testing results more accurate.

However, the difference between the two overall error rates is small (0.2%), which may be due to the number of training sessions, number of iterations, final convergence and small differences in only one set of random values in the convolutional neural network. As observed above, a mixed training dataset yields better detection performance than using a single dataset for training. For example, the detection with the mixed (Carioca 1 and Spanish Speech) training dataset decreases the overall error rate by 1.3% compared with the single Carioca 1 training dataset and 1.1% compared with the single Spanish Speech dataset. The key insight is that the increased data diversity provided by mixed training increases the generality of the learned model, which improves the detection performance and reduces the overall error rate.

Figure 4 compares the detection performance of the proposed scheme using the convolutional neural network with MUSIC, Hilbert linear prediction and Welch features (CNN with MHPW) versus the detection performance of the benchmark strategy by Reis et al. [19]. The same Carioca 1 and Spanish Speech databases with $N = 400$ audio record-

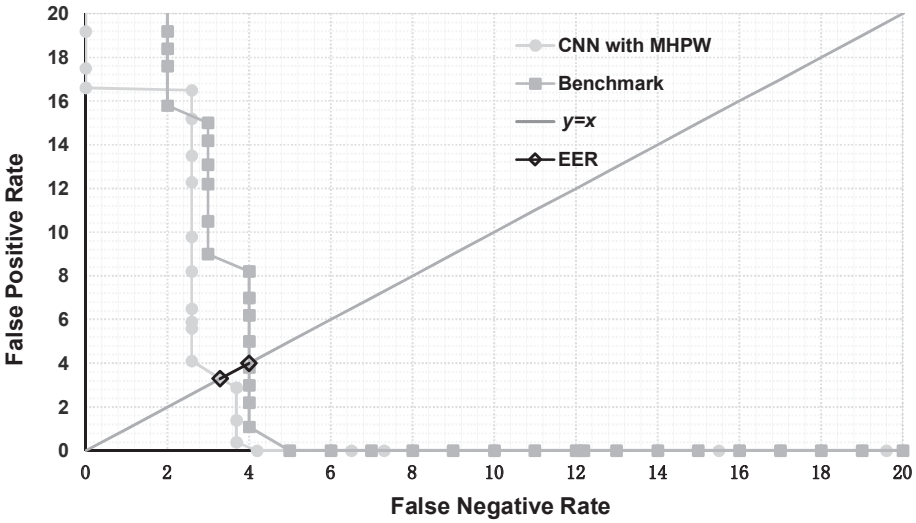

Figure 4. Comparison of detection performance.

ings and $l = 1$ second were used in the comparison. The DET curve of the proposed scheme (CNN with MHPW) is much closer to the y-axis (i.e., lower false negative rates) compared with the benchmark strategy. Moreover, the proposed scheme has an equal error rate of 3.3%, which is less than the 4% equal error rate of the benchmark strategy.

5.3 Results for Different SNR Conditions

This section evaluates the performance of the proposed audio forensic scheme under signal-to-noise ratios ranging from 5 dB to 30 dB. The speech activity detector of Esquef et al. [3] was used to separate the noise from speech signals in the Carioca 1 and Spanish Speech databases, following which various levels of additional background Gaussian noise were introduced. It is important to note that the results of this evaluation can be generalized to any audio recording.

Figure 5 shows the performance of the proposed audio forensic scheme (CNN with MHPW) under various signal-to-noise ratios. The equal error rates obtained for the datasets corrupted by Gaussian noise decrease with increasing signal-to-noise ratio because the classifier acquires more accurate electric network frequency information with less noise. The performance gap is much wider at lower signal-to-noise ratios, which validates the effectiveness of low-rank filtering for noise correction. For example, the proposed scheme decreases the equal error rate from 20.8% to 17.9% for the lowest signal-to-noise ratio of 5 dB. In fact, the proposed

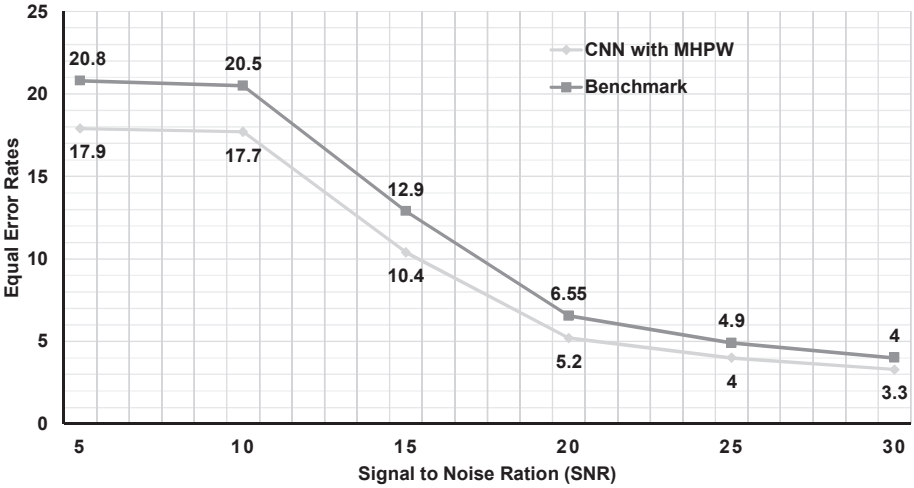

Figure 5. Performance of the audio forensic scheme under different SNR conditions.

scheme achieves optimal performance faster than the benchmark strategy with less signal information. Specifically, the equal error rate of the proposed scheme decreases from 17.9% at 5 dB to 4.0% at 25 dB whereas the benchmark strategy achieves the same equal error rate only at 30 dB. Moreover, despite showing a consistent performance trend, the proposed scheme is more effective than the benchmark strategy even for low signal-to-noise ratios. This demonstrates that the low-rank filtering incorporated in the proposed scheme improves the accuracy of detecting audio tampering, especially in poor signal-to-noise conditions.

6. Conclusions

The audio forensic scheme described in this chapter leverages a convolutional neural network classifier to evaluate electric network frequency features in audio signals to detect tampering without manual regulation or information about the concurrent reference frequency from the power grid. The experimental results demonstrate that the audio forensic scheme increases the accuracy of tamper detection and is better adapted to noisy environments than the benchmark strategy of Reis et al. [19]. For example, the proposed scheme reduces the overall error rate by 1.3% and increases the equal error rate by 0.7% compared with the benchmark strategy. Additionally, it increases the equal error rate up to 2.9% compared with the benchmark strategy under different signal-to-noise conditions. The tamper detection performance and robustness

in noisy environments help ensure the reliability of audio evidence and protect intellectual property.

Future research will attempt to enhance detection accuracy and efficiency in more aggressive scenarios, and develop an online detection system that identifies the specific locations of audio tampering. Additionally, future research will explore the application of electric network frequency signals in video forensics.

Acknowledgement

This research was supported by the Natural Science Foundation of China under Grant nos. 61772571, U1536204, 61971366 and 61671396.

References

[1] M. Arnold, Audio watermarking: Features, applications and algorithms, *Proceedings of the IEEE International Conference on Multimedia and Exposition – Latest Advances in the Fast-Changing World of Multimedia*, vol. 2, pp. 1013–1016, 2000.

[2] J. Chen, X. Kang, Y. Liu and Z. Wang, Median filtering forensics based on convolutional neural networks, *IEEE Signal Processing Letters*, vol. 22(11), pp. 1849–1853, 2015.

[3] P. Esquef, J. Apolinario and L. Biscainho, Edit detection in speech recordings via instantaneous electric network frequency variations, *IEEE Transactions on Information Forensics and Security*, vol. 9(12), pp. 2314–2326, 2014.

[4] P. Esquef, J. Apolinario and L. Biscainho, Improved edit detection in speech via ENF patterns, *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2015.

[5] R. Garg, A. Varna, A. Hajj-Ahmad and M. Wu, "Seeing" ENF: Power-signature-based timestamps for digital multimedia via optical sensing and signal processing, *IEEE Transactions on Information Forensics and Security*, vol. 8(9), pp. 1417–1432, 2013.

[6] A. Hajj-Ahmad, R. Garg and M. Wu, Instantaneous frequency estimation and localization for ENF signals, *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012.

[7] A. Hajj-Ahmad, R. Garg and M. Wu, ENF-based region-of-recording identification for media signals, *IEEE Transactions on Information Forensics and Security*, vol. 10(6), pp. 1125–1136, 2015.

[8] G. Hua, G. Bi and V. Thing, On practical issues of electric network frequency based audio forensics, *IEEE Access*, vol. 5, pp. 20640–20651, 2017.

[9] G. Hua, Y. Zhang, J. Goh and V. Thing, Audio authentication by exploring the absolute error map of ENF signals, *IEEE Transactions on Information Forensics and Security*, vol. 11(5), pp. 1003–1016, 2016.

[10] M. Imran, Z. Ali, S. Bakhsh and S. Akram, Blind detection of copy-move forgery in digital audio forensics, *IEEE Access*, vol. 5, pp. 12843–12855, 2017.

[11] G. Karantaidis and C. Kotropoulos, Assessing spectral estimation methods for electric network frequency extraction, *Proceedings of the Twenty-Second Pan-Hellenic Conference on Informatics*, pp. 202–207, 2018.

[12] X. Lin and X. Kang, Robust electric network frequency estimation with rank reduction and linear prediction, *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 14(4), article no. 84, 2018.

[13] X. Lin, J. Liu and X. Kang, Audio recapture detection with convolutional neural networks, *IEEE Transactions on Multimedia*, vol. 18(8), pp. 1480–1487, 2016.

[14] Y. Liu, Z. Yuan, P. Markham, R. Conners and Y. Liu, Application of power system frequency for digital audio authentication, *IEEE Transactions on Power Delivery*, vol. 27(4), pp. 1820–1828, 2012.

[15] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, The DET curve in assessments of detection task performance, *Proceedings of the Fifth European Conference on Speech Communication and Technology*, 1997.

[16] D. Nicolalde Rodriguez and J. Apolinario, Evaluating digital audio authenticity with spectral distances and ENF phase change, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1417–1420, 2009.

[17] D. Nicolaide Rodriguez, J. Apolinario and L. Biscainho, Audio authenticity: Detecting ENF discontinuity with high precision phase analysis, *IEEE Transactions on Information Forensics and Security*, vol. 5(3), pp. 534–543, 2010.

[18] X. Pan, X. Zhang and S. Lyu, Detecting splicing in digital audios using local noise level estimation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1841–1844, 2012.

[19] P. Reis, J. da Costa, R. Miranda and G. Del Galdo, ESPRIT-Hilbert-based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency, *IEEE Transactions on Information Forensics and Security*, vol. 12(4), pp. 853–864, 2016.

[20] H. Su, A. Hajj-Ahmad, M. Wu and D. Oard, Exploring the use of ENF for multimedia synchronization, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4613–4617, 2014.

[21] S. Vatansever, A. Dirik and N. Memon, Factors affecting ENF-based time-of-recording estimation for video, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2497–2501, 2019.

[22] Z. Wang, J. Wang, C. Zeng, Q. Min, Y. Tian and M. Zuo, Digital audio tampering detection based on ENF consistency, *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, pp. 209–214, 2018.

[23] Q. Zhao, D. Meng, Z. Xu, W. Zuo and L. Zhang, Robust principal component analysis with complex noise, *Proceedings of the Thirty-First International Conference on Machine Learning*, vol. II, pp. 55–63, 2014.