SUPERVISED AUDIO TAMPERING DETECTION USING AN AUTOREGRESSIVE MODEL

Xiaodan Lin[1,2], *Xiangui Kang*[1]

1. School of Data and Computer Science, Sun Yat-Sen University, 510006, Guangzhou, China
2. School of Information Science and Engineering, Huaqiao University, Xiamen 361021, China

ABSTRACT

Splicing, cutting and insertion are the most common operations imposed on audio files when the adversary intends to modify or fabricate the content. The detection of such kinds of tampering is still challenging in real-world applications. In this paper, a generic approach for the detection of audio tampering is proposed via the analysis of electric network frequency (ENF). Based on the fact that tampering with an audio leads to anomalous variations of the underlying ENF signal, a wavelet-filtered ENF signal is generated to highlight the abnormal ENF variations. An autoregressive (AR) model is then fitted to the detail part of the ENF signal and the resulting AR coefficients are employed to train the classifier under a supervised-learning framework. Experimental results show that our proposed method significantly outperforms the state-of-art methods in the context where moderate or high levels of noise are present. Moreover, robustness against MP3 compression can be achieved.

Index Terms— Audio tampering detection, electric network frequency, supervised learning, autoregressive model

1. INTRODUCTION

Over the past decade, digital audio files have proliferated on the Internet and in every aspect of our daily lives. If the audio content is distorted for some illegal purposes and further distributed via networks, or presented as evidences to the court, severe social problems may arise. Therefore, the identification of forged audio recordings has become an essential task in the forensic society.

The widely used techniques that can modify the audio contents are splicing, copy-move and cutting. In the two latter cases, no extrinsic audio sources are added in. Audio tampering artifacts are usually imperceptible for human hearing, so it is necessary to design a detector that can automatically expose these artifacts. Countermeasures that are devised to reveal traces left by audio tampering include the local noise level estimation method [1], which assumes that the splicing audio contains different noise levels. However, this does not hold true for the cases of copy-move and cutting since the noise level of the tampered audio remains consistent. By exploring pitch similarity, a detector presented in [2] is developed specifically for the copy-move forgery. The authors in [3] first report that the electric network frequency (ENF), which randomly deviates from the nominal frequency of 50 Hz or 60 Hz, can be captured in the audio recordings. From then on, multiple studies have lent this finding to evaluate audio authenticity [4-5]. In [6], an efficient method to authenticate audio signals is proposed by detecting ENF phase discontinuity. This approach is further extended in [7] to show that higher harmonics of an ENF signal can also provide evidence of audio tampering. A recent work operated on ENF abnormality is reported in [8], where the authors employ a data-driven threshold-based strategy to deal with the anomalous variations of the ENF signal. An improvement has been made in [9] by taking the patterns of ENF variation into account. Although the methods in [8-9] have provided superiority in terms of detection accuracy to the methods in [6-7], they still suffer greatly from the interference of potential noise. Besides, whether these detectors can survive transcoding artifacts like MP3 compression has not been investigated.

To overcome these limitations, we focus on designing a more robust detector to expose audio tampering. Triggered by the work in [8], we observe that not only the ENF phase, but also the detailed ENF fluctuations can reveal the anomalous traces left by audio tampering. Upon appropriate modeling of the detailed ENF fluctuations, we propose a novel method to detect audio forgery from a supervised-learning perspective, rather than the standard signal processing techniques employed by the existing detectors. The benefit is twofold. First, it incorporates all the possible audio tampering manipulations, therefore a general framework for audio tampering detection can be derived. Second, unlike the existing methods, we don't have to manually tune the thresholds or parameters in order to gain an appropriate detector. To achieve this, autoregressive modeling is employed to yield compact but robust features.

Corresponding Author: isskxg@mail.sysu.edu.cn. This work was supported by NSFC (Grant nos. 61379155, U1536204), NSF of Guangdong province (Grant no. s2013020012788) and Research Fund of Fujian Educational Committee (Grant no. JAT160036).

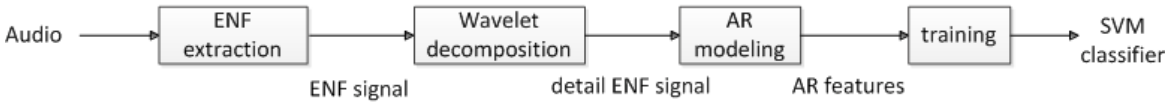

Fig. 1. A block diagram of the proposed audio tampering detection system.

2. AN OVERVIEW OF THE SYSTEM

An overview of the proposed system for audio tampering detection is shown in Figure 1. First, the underlying ENF signals are extracted from all the speech signals in the training set. Second, each of the extracted ENF signals is exposed to a wavelet-filtering process, yielding the detail part of the ENF signal, which is then modeled by an autoregressive process. The resulting AR coefficients are used as input features to train the SVM classifier. In the testing stage, the AR coefficients are obtained from a test audio by following the same paradigm in the training stage and are adopted as the input to the classifier. The label given by the classifier indicates whether the audio is a forged one or not. For example, "1" denotes a forged audio while "0" denotes that the audio is genuine. It should be pointed out that the training audio data incorporated in this work are in uncompressed waveform, whereas the test audio can be in forms of either uncompressed PCM or compressed MP3 with various bitrates.

3. DETECTION OF AUDIO TAMPERING

3.1. ENF Extraction

Prior to the feature engineering process, the ENF signals should be first extracted from the given audio recordings. Several techniques have been developed for the extraction of ENF signals [10-13]. Since the aim of our proposed scheme is to capture the abnormal ENF variations, rather than the fine-grained matching of the extracted ENF to the referenced ENF in applications like timestamp verification where more accurate ENF tracking is required, we employ a computationally simple approach for ENF extraction, i.e., the weighted spectrogram-based approach [12]. To be more specific, the audio signal is downsampled and band-pass filtered so as to concentrate on the frequency bands of our interest. Due to the fact that ENF signals fluctuate slowly around the nominal frequency of 50 Hz or 60 Hz, band-pass filter centered on the nominal frequency is employed. The ENF signal $x(n)$ is calculated by weighting the frequency of each short-term frame according to (1).

$$x(n) = \frac{\sum_{l=L_1}^{L_2} f(n,l)|S(n,l)|}{\sum_{l=L_1}^{L_2} |S(n,l)|} \tag{1}$$

where $L_1 = \lfloor (f_0 - 0.5)N/f_s \rfloor$ and $L_2 = \lceil (f_0 + 0.5)N/f_s \rceil$; f_0, f_s and N are the nominal ENF frequency, the sampling frequency and the number of FFT points. $f(n,l)$ and $|S(n,l)|$ denote the frequency and energy in the *l*-th frequency bin of the *n*-th short-term analysis frame. We use a frame length of 1 second and an overlap factor of 0.9, indicating that ten ENF estimations are obtained every second. An FFT length of N is fixed at 4096 using zero-padding and the sampling frequency is 500 Hz. Figure 2 shows an example of audio tampering, in which a segment of the audio is cropped away from the original audio. It can be observed in Figure 2 that for the original audio, the ENF variation shows to be more stationary; while for the tampered audio, abrupt changes of the extracted ENF can be seen. Considering that the length of the extracted ENF differs due to the varying lengths of audio recordings, the ENF signal itself cannot be readily applied to a supervised classification framework. Therefore, the variation of the ENF needs to be further explored to make it more generic for different types of forgery and different lengths of audios.

3.2. AR Modeling of the Detail ENF

In order to construct features that can effectively capture the inherent correlations of the extracted ENF and be immune to distortions caused by noise or transcoding, an effective modeling of the ENF signal is the essence. It should be noticed that it has already been pointed out in [14] that a first-order autoregressive process can be applied to model the ENF signals and has proved to be useful for timestamp verification of audio recordings.

Considering that the abnormal fluctuations of the ENF signal are more prominent in the high-frequency part, the autoregressive modeling is not straightly enforced. Instead, wavelet decomposition is performed, such that the detail ENF fluctuation can be obtained. The detail ENF signal $d(n)$ can be formulated by

$$d(n) = H[x(n)] \tag{2}$$

where $x(n)$ is the extracted ENF signal and $H[\cdot]$ denotes a filtering operation. It can be found in Figure 2(d) that the extracted ENF signal varies smoothly in the regions where tampering does not occur, whereas a burst shows up at the borders of tampering, similar to impulse noises. Inspired by this finding, wavelet decomposition [15] is employed to present fine-grain details of the ENF fluctuation. It should be noted that one-dimensional wavelet decomposition is sufficient to handle the ENF signal. The detail ENF signals for the original

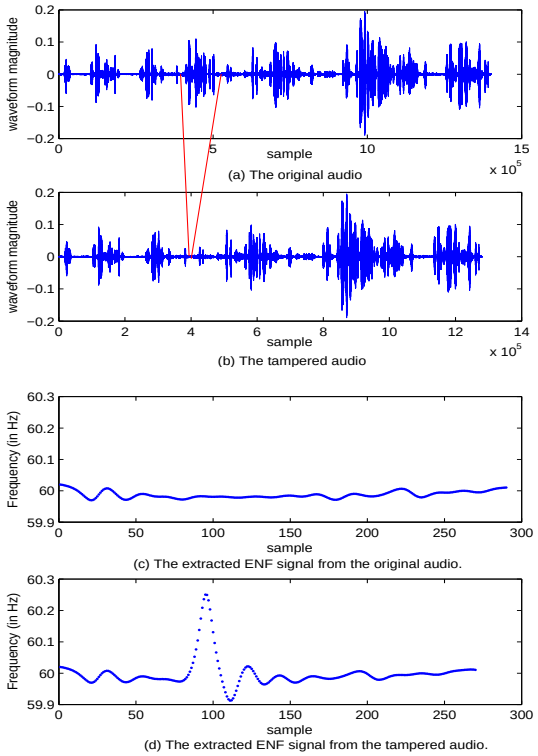

Fig. 2. An example of audio tampering.

audio and its tampered counterpart are shown in Figure 3. It is found that the detail ENF of the untampered audio fluctuates around zero and shows high correlations over time, whereas the abnormal variation caused by tampering undermines this correlationship. Motivated by this result, the autoregressive model [16] which well depicts correlated time series is applied to construct the features for further supervised classification. An m-th order AR process involved in this work is described as

$$d(n) = \sum_{i=1}^{m} a_i d(n-i) + e(n) \tag{3}$$

where a_i denotes the AR coefficient and $e(n)$ denotes the prediction error. The total m-th order AR coefficients are estimated via the Burg method and used as features for training and classification.

4. EXPERIMENT

To validate our proposed scheme, a set of experiments are carried out and the classification performance is given in this section. Further, the effect of varying levels of noise is discussed. Finally, the effectiveness of the proposed scheme under MP3 compression will be studied.

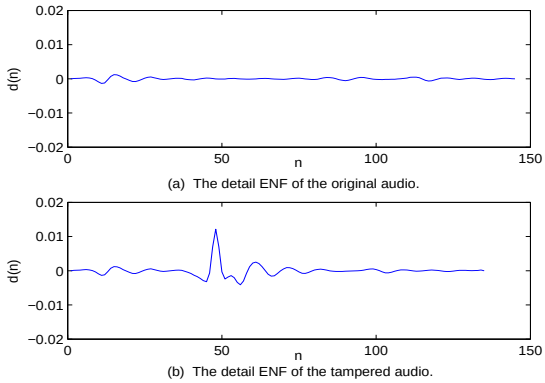

Fig. 3. The detail ENF signal.

4.1. Experimental Setup

To facilitate the comparison with the existing method, we employed the same speech database as [9] for a thorough assessment. The Carioca 1 database with a sampling rate of 44.1 kHz embraces an ENF component around 60 Hz while the Spanish Speech Database sampled at 16 kHz contains an ENF component around 50 Hz. Each of the two databases consists of 200 speech signals, half of which are original while the other half is tampered with by means of deletion or insertion (copy-move). For each of the tampered speech signals, only one edit is made and the edit position is chosen at voice-inactive parts of the signal.

In training an SVM classifier, the RBF kernel is employed to map the input features to a higher dimensional space so as to make them distinguishable [17]. Five-fold cross validation is conducted to choose the kernel parameters. According to Box-Jenkins methodology [18], the order of AR coefficients is set to 14 in the experiments.

4.2. Detection Performance

First, the classifier is trained using features from 160 randomly selected speech signals from the two databases and testing is conducted on the remaining 240 speech signals. The original audio recordings and tampered audio recordings are of equal size for both the training set and the testing set. Table 1 shows the confusion matrix. The detection accuracy is 97.5%, with a true positive rate of 99.2% and a true negative rate of 95.8%, indicating that the proposed AR features own better discriminative capability for the tampered audio files.

Second, we conduct a cross domain evaluation based on the two databases described above, i.e., the Carioca 1 database is used as the training set and the Spanish Speech Databases as the testing set, and vice versa. The detection accuracy is shown in Table 2. It is observed that the AR features trained on the Carioca 1 dataset deliver slightly better performance than the features learned on the Spanish dataset. Neverthe-

Table 1. Confusion matrix of the proposed detector

	Classified		
Ground-truth		**original**	**tampered**
	original	95.8%	4.2%
	tampered	0.8%	99.2%

Table 2. Cross-domain evaluation

Test Domain	Features	Accuracy
Carioca 1	Learned on Spanish	96.3%
Spanish	Learned on Carioca 1	96.9%

less, features can be effectively learned from either dataset even if they possess distinct waveform parameters and SNR conditions. Compared with the results in Table 1 where training data are a mixture of audio signals from both databases, lower performance is observed in the cross domain evaluation. An insight gained from this performance gap is that by increasing the diversity of training samples, i.e. diverse patterns of ENF variation, the classification performance can be boosted.

4.3. The Effect of Noise

It should be noted that in real-life applications, the noise condition can be far worse. In the following, we will show how the proposed detector reacts to various noise conditions. Various levels of white Gaussian noise are considered and are added to the Spanish Speech dataset with the same sampling rate. In this test, the features are learned from the clean Carioca 1 dataset and tested on the noisy Spanish Speech dataset. To render a fair comparison with the method in [9], the performance of the proposed binary classifier is measured in terms of the equal error rate (EER), where the probability of false positive is equal to the probability of false negative. The overall detection performance in terms of EER is reported in Table 3. As shown in Table 3, our proposed method significantly outperforms the method in [9] by 8.4% when the SNR is set to 20dB. When the noise condition gets worse, the performance gap between the two methods increases.

Table 3. Assessment under various noise levels in EER

SNR (in dB)	Ours	method in [9]
25	2.9%	3.2%
20	4.6%	13%
15	11%	41.4%
10	12.5%	47%
5	14.5%	48.5%

Table 4. Assessment under MP3 compression in EER

Bit rate	Ours	method in [9]
64 kbps	8%	32%
128 kbps	6.4%	15.5%
192 kbps	3.6%	14%

4.4. Robustness against MP3 Compression

To evaluate the robustness of our proposed scheme against MP3 compression, the Spanish dataset is compressed with bitrates of 64 kbps, 128 kbps and 192 kbps. Cross-domain evaluation is performed in this test, i.e., the features are learned on the uncompressed Carioca 1 dataset and tested on the compressed Spanish dataset with the above three bitrates. The detection results are reported in Table 4. As seen in Table 4, even if the audio is compressed at a bit rate as low as 64 kbps, our proposed scheme can still distinguish both the original and tampered audio with an EER of 8%, which demonstrates the robustness of the proposed method to MP3 compression. In contrast, the method in [9] is far more susceptible to MP3 compression, which may be accounted by fixed editing templates assumed therein. To be more specific, a potential mismatch between the ENF variation and the assumed templates can be introduced due to the MP3 compression. Besides, the ENF estimation via Hilbert's method adopted in [9] is more sensitive to MP3 compression especially when the bit rate is low, leading to inaccurate ENF estimation.

5. CONCLUSIONS

In this work, an effective detector for multiple types of audio forgeries is developed. Based on the fact that tampering with an audio leads to anomalous variations of the underlying ENF signal, we apply a wavelet filter to the extracted ENF, followed by an autoregressive modeling of the detail ENF signal. A supervised classification framework which exploits the statistical autoregressive features is introduced to identify whether an audio is a tampered one or not. Compared to the existing works, the advantages of our proposed method lies mainly in three aspects. First, our proposed method achieves significant improvements in noisy conditions and provides robustness against MP3 compression. Second, the AR features used as inputs to the classifier greatly reduce the feature dimension, as it is an important concern in machine learning applications. Third, upon the supervised learning framework, a decision can be automatically rendered without the need of manually tuning the parameters or the thresholds. It is worth noting that our method can directly apply to audios of short durations, for example, audio length used in this study ranges from 19 seconds to 39 seconds. When our proposed method is applied to a longer audio, a segment-wise examination should be utilized. This is because the ENF signal assumes to be piecewise stationary, under which the AR modeling works.

6. REFERENCES

[1] X. Pan, X. Zhang, S. Lyu, "Detecting splicing in digital audios using local noise level estimation," In *Proc. of ICASSP*, pp. 1841-1844, Kyoto, Japan, March 2012.

[2] Qi Yan, Rui Yang, Jiwu Huang, "Copy-move detection of audio recording with pitch similarity," In *Proc. of ICASSP*, pp. 1782-1786, Brisbane, Australia, April 2015.

[3] C. Grigoras, "Digital audio recordings analysis: The electric network frequency (ENF) criterion," *Int. J. Speech Lang. Law*, vol. 12, no. 1, pp. 63-76, 2005.

[4] R. W. Sanders, "Digital authenticity using the electric network frequency," In *Proc. 33rd AES Int. Conf. Audio Forensics, Theory Practice*, pp. 1-6, Jun. 2008.

[5] R. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 84-94, 2009.

[6] D. 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, no. 3, pp. 534-543, Sept. 2010.

[7] D. Nicolalde, J. Apolinario Jr., and L. Biscainho, "Audio authenticity based on the discontinuity of ENF higher harmonics," In *Proceedings of the 21st European Signal Processing Conference*, pp. 1-5, Marrakech, Maroc, Sept. 2013.

[8] P. Esquef, J. Apolinario Jr., and L. Biscainho, "Edit Detection in Speech Recordings via Instantaneous Electric Network Frequency Variations," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2314-2326., Dec. 2014.

[9] P. Esquef, J. Apolinario Jr., and L. Biscainho, "Improved edit detection in speech via ENF patterns," In *IEEE International Workshop on Information Forensics and Security*, pp.1-6, Rome, Italy, Nov. 2015.

[10] O. Ojowu, J. Karlsson, J. Li, and Y. Liu, "ENF extraction from digital recordings using adaptive techniques and frequency tracking," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1330-1338, Aug. 2012.

[11] A. Hajj-Ahmad, R. Garg, and M. Wu, "Spectrum combining for ENF signal estimation," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 885-888, Sep. 2013.

[12] R. Garg, A. L. Varna, and M. Wu, "'Seeing' ENF: natural time stamp for digital video via optical sensing and signal processing," In *Proc. of the 19th ACM Intl. conf. on Multimedia*, pp. 23-32, Scottsdale, AZ, USA, 2011.

[13] L. Fu, P. Markham, R. Conners and Y. Liu, "An Improved Discrete Fourier Transform-Based Algorithm for Electric Network Frequency Extraction," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1173-1181, July. 2013.

[14] R. Garg, A. Varna, and M. Wu, "Modeling and analysis of electric network frequency signal for timestamp verification," In *Proc. of IEEE Workshop on Info. Forensics and Security*, pp. 67-72, Tenerife, Spain, Dec. 2012.

[15] A. Abdelnour and I. Selesnick, "Nearly symmetric orthogonal wavelet bases," In *Proc. of ICASSP*, May 2001.

[16] S. Haykin, Adaptive Filter Theory, Prentice-Hall, Inc., 2001.

[17] C. - C. Chang and C. - J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol.2, no.3, 389-396, Apr. 2011.

[18] G. Box, G. Jenkins, and G. Reinsel, "Time Series Analysis: Forecasting and Control," Hoboken, NJ, USA: Wiley, 2008.