AUDIO FORENSIC AUTHENTICATION BASED ON MOCC BETWEEN ENF AND REFERENCE SIGNALS

Zhisheng Lv[1], Yongjian Hu[1], Chang-Tsun Li[2], and Bei-bei Liu[1]

[1]School of Electronic and Information Engineering,
South China University of Technology, Guangzhou 510640, P.R.China
[2]Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

ABSTRACT

This paper proposes a new audio authenticity detection algorithm based on the max offset for cross correlation (MOCC) between the extracted ENF (Electric Network Frequency) signal and the reference signal. We first extract the ENF signal from a query audio signal. And then we partition it into overlapping blocks for forgery detection. The MOCC between the extracted ENF and the reference signal is calculated block by block. We also introduce an enhancement scheme to improve the quality of the ENF signal before the calculation of the MOCC. Our proposed method can detect not only audio forgery but also the edited region and audio forgery type. The effectiveness of our method has been verified by experiments on digitally edited audio signals.

Index Terms—Electric network frequency; audio forgery detection; max offset for cross correlation; single-frequency reference signal.

1. INTRODUCTION

Digital audio recording is much convenient nowadays and audio editing also becomes a simple task by using modern audio editing software. Even non-professionals can modify audio without leaving any visible traces [1]. Therefore, the detection of digital audio authenticity and integrity becomes significant. Some progress has been made in the past few years, e.g., [2] and [3]. One way of audio forgery detection is to use the electric network frequency (ENF). When recording with digital equipment connected to the power grid, the ENF signal is inevitably embedded in the recorded audio signal. By examining the change of ENF signal, the authenticity of a query audio can be verified. In [5]-[8], the ENF pattern extracted from a query recording was compared with the recorded historical ENF signal database to determine continuity and consistency of the audio. In [9], Nicolalde *et al.* presented a method to identify audio authenticity by analyzing the change of phase of the ENF signal. In [10] the same authors proposed an improved work and used a high-precision Fourier analysis method to better estimate the ENF phases for audio forgery detection. The method of using the ENF for forensic investigation was also extended to video (e.g., [4]). Motivated by [9] and [10], we propose a new audio forgery detection method in this work.

2. ALGORITHM PRINCIPLE AND IMPLEMENTATION

For simplicity of description, we first define the notations and symbols employed in Table 1.

Table 1 Notations and symbols.

s_{ENF}	ENF signal
d	narrow band noise
x	extracted ENF signal
f_0	the nominal ENF
f_s	sample frequency
U_r	amplitude of reference signal
N	window size for enhancement
N_C	window size for the max offset
N_{DFT}	points of DFT operation
N_0	number of samples per cycle of the nominal ENF ($N_0=f_s/f_0$)
K	number of blocks for the max offset
n	index of signal samples
i	index of samples in every block, $i=1,2,\cdots N_C$
k	index of blocks, $k=1,2,\cdots K$

2.1. Algorithm principle

For the power grid, the fluctuation of ENF signal over time is strictly controlled [1][6]. In other words, the initial phase of ENF signal in every cycle is consistent for an unedited original audio. If the audio signal has undergone forgery, the initial phase of its ENF signal will change abruptly at the edited point [9]. Based on the consistency of phase of ENF signal in audio, the authors of [10] proposed a way to detect audio edition. They relied on the high-precision Fourier analysis method to estimate the phase of ENF signal. They first divided the ENF signal into overlapping blocks. In practice, they used a sliding window that covers multiple cycles of the ENF signal. And then they determined audio authenticity by checking the change of phase. One obvious

weakness of their method is high computational complexity due to the DFT (discrete Fourier transform). Another problem is that the forgery boundary detected is not very accurate.

This work proposes a more efficient and accurate method. We also divide the query audio signal into overlapping blocks like [10]. Then for every block we determine the offset for the instance where the cross correlation between the extracted ENF signal and the reference signal reaches the maximum. We call this offset the *max offset* (i.e., MOCC) for simplicity. The MOCC is closely related to the phase of ENF signal. If the phase is consistent for every block of the query signal, the MOCCs are the same for all blocks; otherwise, the MOCCs would be different. Hence, we can determine the audio authenticity by checking the change of MOCCs from different blocks. The calculation of MOCC is a spatial domain operation, and does not need the DFT which was used by [10], so our method has lighter computational load.

2.2. Calculation of MOCC between the extracted ENF signal and the reference signal

In general, the ENF signal extracted by the narrow band-pass filter can be expressed as follows

$$x(n) = s_{ENF}(n) + d(n) \tag{1}$$

where $s_{ENF}(n) = A_0 \cos[2\pi n(f_0 + \xi(n))/f_s + \theta]$, A_0, f_0, θ refer to the amplitude, designed frequency and initial phase of ENF signal, respectively. $\xi(n)$ denotes the ENF fluctuation caused by the power grid. $d(n)$ consists of the background noise and the residual speech signal for filter leakage.

To detect audio forgery, we introduce a single-frequency reference signal, which is a cosine signal with the same frequency as the ENF signal

$$r(n) = U_r \cos(2\pi n f_0 / f_s) \tag{2}$$

As mentioned before, we divide $x(n)$ into overlapping blocks by a sliding window of size N_C. Each block overlaps the former by (N_C - N_0) samples. We then calculate the MOCC between $x(n)$ and $r(n)$. For the k-th block, we have

$$\begin{aligned} R_{x,r}(\tau) &= E[x(i)r(i+\tau)] \\ &\approx \frac{U_r A_0}{2}\cos(2\pi\tau\frac{f_0}{f_s} - \theta_k) + \eta(\tau) \end{aligned} \tag{3}$$

where $\tau = 0,1,2,\cdots,N_0 - 1$, and θ_k is the initial phase for the k-th block. τ refers to the offset. The detail of (3) can be seen in APPENDIXES A. Its first item denotes the cross correlation between two cosine signals, and the second refers to the cross correlation between the reference signal and the narrow band noise. If N_C is an integral multiple of N_0 and $\theta_k = 2\pi\tau/N_0$, we obtain

$$\cos(2\pi\tau/N_0 - \theta_k) = 1 \tag{4}$$

This offset value $\tau = N_0\theta_k/2\pi$ is denoted as the max offset $\tau_{\max}[k]$, i.e., the MOCC for the k-th block. Then equation (3) can be rewritten as

$$R_{x,r}(\tau_{\max}[k]) \approx \frac{U_r A_0}{2} + \eta(\tau_{\max}[k]) \tag{5}$$

Due to the finite length of the block, the value of $\eta(\tau_{\max}[k])$ will be close to 0 rather than 0. For an audio signal with a high SNR (signal-to-noise ratio), we have $A_0 \gg |\eta(\tau_{\max}[k])|$. So the noise item has little impact on $R_{x,r}(\tau_{\max}[k])$. Note that we have no interest in calculating $R_{x,r}(\tau_{\max}[k])$ but $\tau_{\max}[k]$ for the purpose of forgery detection. To determine it, we increase τ by a tiny increment each time, and the largest correlation value corresponds to $\tau_{\max}[k]$. $\tau_{\max}[k]$ is the nearest integer of $N_0\theta_k/2\pi$ and is irrelevant to N_C. If the MOCC values are same or nearly the same for all the blocks of $x(n)$, the audio is original or has not undergone edition; otherwise, it has been edited.

2.3. Enhancement scheme

In order to suppress the interference of $\eta(\tau)$ in (3), we propose an enhancement scheme. Let us take the output of (3) as the new input and carry out (3) again. From the computation in APPENDIXES B, we can get

$$R_{x,r^{(2)}}(m) \approx (\frac{U_r}{2})^2 A_0 \cos(2\pi m\, f_0/f_s + \theta_k) + \eta^{(2)}(m) \tag{6}$$

where m=1,2,…,N. The superscript of η refers to the iteration times. Let U_r =2. The first item of (6) looks like that in (1) if we ignore $\xi(n)$. The difference is the second item which reflects the noise component. We can easily find that $\eta^{(2)}$ is less than $\eta^{(1)}$ (i.e., η in (3)) and $\eta^{(1)}$ is less than $d(n)$. Repeat this process 2M times, we get

$$R_{x,r^{(2M)}}(m) \approx A_0 \cos(2\pi m\, f_0/f_s + \theta_k) + \eta^{(2M)}(m) \tag{7}$$

where m=1,2,…,N. The first item of (7) is similar to that of (1) while the second item is much smaller than $d(n)$. So we use the output of (7) to replace $x(n)$ in (3) when we calculate the MOCC for each block. Apparently, this process is similar to a signal enhancement operation. Our enhancement scheme is motivated by the multilayer autocorrelation in [11]. To avoid high computational complexity, the enhancement process is limited within a sliding window of size N. N should be larger than N_C. Since larger M may cost more computation time, we let M=2, which yields the good effect empirically. We will use experiments to demonstrate the effect of our enhancement scheme later.

2.4. Algorithm implementation

The difference of the MOCCs for adjacent blocks can be expressed as $\Delta\tau_{\max}[k]=\tau_{\max}[k+1]-\tau_{\max}[k]$. We propose to use $\Delta\tau_{\max}$ for audio authenticity detection. The steps of algorithm implementation are as follows: (a) Down-sample query audio to 1000Hz or 1200Hz to reduce computational cost; (b) Extract the ENF signal by a linear-phase band-pass filter; (c) Enhance the ENF signal using (3); (d) Calculate the MOCC; (e) Determine the audio forgery by $\Delta\tau_{\max}$; (f) Evaluate the edited region and forgery type.

3. EXPERIMENT AND ANALYSIS

We compare the proposed method with the method in [10]. The signals employed for experiments are derived from two public databases, AHUMADA and GAUDI [12], which contain 100 pieces of audio. The nominal ENF of all audio signals is 50 Hz. So we let f_s=1000Hz [10]. We set an audio database that contains the whole 100 original audio files (same as [10]), and 130 intentionally edited audio files (edited by ourselves). Half of the 130 edited audio recordings have an audio portion deleted, while the rest have a portion of audio inserted. In order to avoid strong short-time spectral changes that may make the detection easier, the inserted fragments come from the same file of the edited audio.

3.1. Effect of our enhancement

For a randomly selected audio signal from the database, we add white Gaussian noise WGN(0,σ) into it to generate its noisy version. With the change of σ, we can produce the noisy version of different SNRs. Fig.1 shows the results of comparison. The SNR1 corresponds to the SNR of audio signal while the SNR2 describes the SNR of the non-enhanced ENF signal and the enhanced one. The non-enhanced ENF signal is (1) while the enhanced one refers to the output of (7). The SNR2 for the enhanced one is apparently higher. Furthermore, larger *N* would yield better results.

3.2. Detection of deletion

We randomly choose an edited audio signal from the database, which has suffered from removing the fragment between 5.48s and 7.72s. Fig. 2 shows the result of deletion detection. In Fig.2, we can see that both the method in [10] and our method can detect the forgery, but the latter is more accurate for the determination of forgery boundary due to the sharp change of MOCCs. Note that there is a false alarm which occurs near cycle 375. It can be removed by using a suitable threshold. In general, we can set two thresholds, one for controlling the magnitude of differences of the MOCCs, the other for controlling the interval of the two blocks for

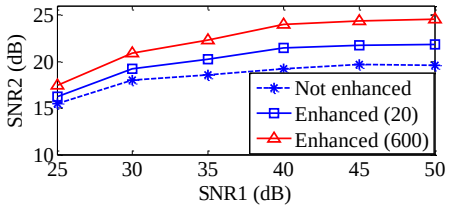
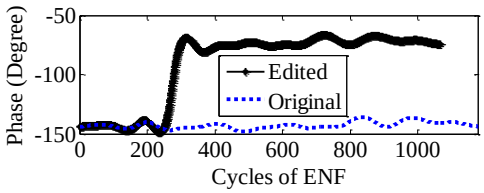
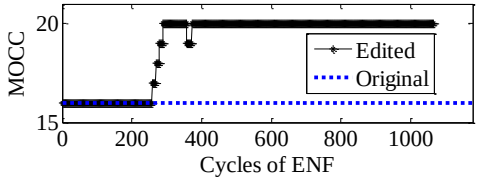


Fig. 1. Enhancement results with *N*=20 and *N*=600, respectively.



(a) Estimated phases using the method in [9]



(b) MOCCs

Fig. 2. Results of deletion detection.

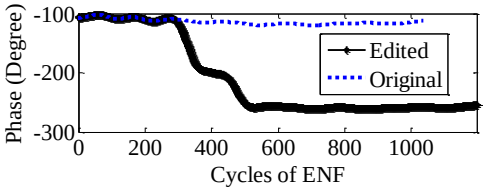
comparison. Due to the limitation of paper length, we cannot discuss this problem further. With $N_0=f_s/f_0$=20, the index of the forgery boundary (i.e., *k*) and the period of each cycle (i.e., 0.02s), the edited region can be automatically determined in the interval of 5.41s and 5.53s, which is very close to the truly deleted one (i.e., 5.48s). In contrast, the method in [10] cannot obtain such a precise location due to the vague boundary. Since the voice activity is absent in the estimated edited area, our method can judge the forgery type as deletion.

3.3. Detection of insertion

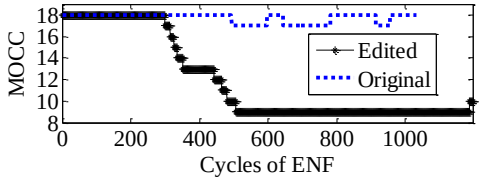
We randomly choose an edited audio signal from the database, in which a 3.27s long segment has been inserted at the instant 6.33s. In Fig.3, both the methods are able to detect the forgery, but our method can locate the starting and ending blocks of the edited region more accurately. Similar to subsection 3.2, the time interval for the edited region can be calculated. It says from 6.31s to 9.55s, which is very close to the real one. From the content of the estimated region, the forgery type can be judged as insertion. In contrast, the vague boundaries prevent the method in [10] from giving the results as accurate as our method.

3.4. Computational complexity

We briefly compare the computational complexity of our method with the method in [10] for the insertion detection in Fig.3. Table 2 indicates the superiority of our method.



(a) Estimated phases using the method in [9]



(b) MOCCs

Fig. 3. Results of insertion detection.

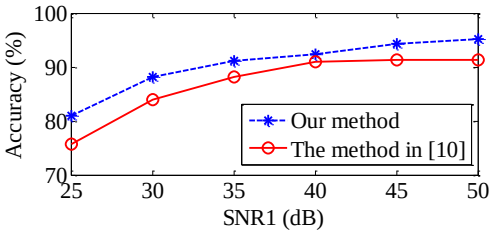


Fig. 4. Accuracy under different SNRs.

Table 2 Comparison of computation time.

Parameter	Method in [10]	Our method
N_{DFT} =500,N_C=60	1.073s	0.659s
N_{DFT} =1000,N_C=100	1.273s	0.661s
N_{DFT} =2000,N_C=200	1.783s	0.662

Moreover, with the increase of N_{DFT} and N_C, the time required for the method in [10] increases much faster than our method. Specifically, the computation time for our method increases very slowly. Similar phenomena can be observed for the deletion detection.

3.5. Robustness against noise

The whole database that contains 100 original audio signals and 130 edited audio signals is used for this experiment. Fig.4 shows the accuracy of the method in [10] and our method with their optimal thresholds under different SNRs. The accuracy is calculated by $AR=(1-(f_p+f_n)/2) \times 100\%$, where f_p is the false alarm rate and f_n the miss detection rate. Apparently, our method has better performance than the method in [10]. Even under SNR=25dB, the accuracy of our method is above 80%. The results prove that our method is more robust against noise than the method in [10].

4. CONCLUSION

We have presented a new method using the MOCC for audio forensic authentication. To suppress the noise component, we have introduced an enhancement scheme for the calculation of the MOCC. By comparing the change of MOCCs from different blocks, we can determine the audio authenticity. Our method can also tell the boundaries of the edited region and forgery types. Since our method is a spatial domain one and does not need to carry out the DFT, the computational load is much lower than the typical method in the literature. Our future work will extend the MOCC-based method to compressed audio signals for forgery detection.

Acknowledgement: This work was supported by the Fundamental Research Funds for the Central Universities, SCUT (Project No. 2012ZM0027). The authors would like to thank Dr. D. P. Nicolalde Rodriguez for offering the source code and audio database for [10].

APPENDIXES

A. THE CROSS CORRELATION BETWEEN $x(n)$ AND $r(n)$

$$R_{x,r}(\tau)=E[x(i)r(i+\tau)]$$

$$=\frac{U_r A_0}{2}\{E[\cos(2\pi(2i+\tau)\frac{f_0}{f_s}+2\pi i\frac{\xi(i)}{f_s}+\theta_k)]+E[\cos(2\pi\tau\frac{f_0}{f_s}$$

$$-2\pi i\frac{\xi(i)}{f_s}-\theta_k)]\}+E[d(i)U_r\cos 2\pi(i+\tau)\frac{f_0}{f_s}]$$

Under the assumption that $d(i)$ and $\cos 2\pi(i+\tau) f_0/f_s$ are independent of each other, and $E[d(i)]=0$, we have $E[d(i)U_r\cos 2\pi(i+\tau) f_0/f_s]=0$ if the sequence is infinite. However, the observation time is limited in real-world scenarios, so $E[d(i)U_r\cos 2\pi(i+\tau) f_0/f_s]$ is close to but not 0. For each τ, it has a non-zero value. We use $\eta(\tau)$ to denote this non-zero value. Such a value would decrease with the increase of the sequence length.

For $\xi(i) \ll f_s$, if the discrete sequence length N_C is an integral multiple of N_0, we can get

$$E[\cos(2\pi(2i+\tau)\frac{f_0}{f_s}+2\pi i\frac{\xi(i)}{f_s}+\theta_k)]=\frac{1}{N_C}\sum_{i=1}^{N_C}\cos(2\pi(2i+\tau)\frac{f_0}{f_s}+$$

$$2\pi i\frac{\xi(i)}{f_s}+\theta_k)\approx\frac{1}{N_C}\sum_{i=1}^{N_C}\cos(2\pi(2i+\tau)\frac{f_0}{f_s}+\theta_k)=0$$

$$E[\cos(2\pi\tau\frac{f_0}{f_s}-2\pi\tau\frac{\xi(i)}{f_s}-\theta_k)]=\frac{1}{N_C}\sum_{i=1}^{N_C}\cos(2\pi\tau\frac{f_0}{f_s}-2\pi\tau\frac{\xi(i)}{f_s}-$$

$$\theta_k)\approx\frac{1}{N_C}\sum_{i=1}^{N_C}\cos(2\pi\tau\frac{f_0}{f_s}-\theta_k)=\cos(2\pi\tau\frac{f_0}{f_s}-\theta_k)$$

So we get

$$R_{x,r}(\tau)=E[x(i)r(i+\tau)]\approx\frac{U_r A_0}{2}\cos(2\pi\tau\frac{f_0}{f_s}-\theta_k)+\eta(\tau)$$

B. THE COMPUTATION OF EQUATION (6)

$$R_{x,r^{(2)}}(m) \approx E\{[\frac{U_r A_0}{2}\cos(2\pi j\frac{f_0}{f_s}-\theta_k)+\eta(j)][U_r\cos 2\pi(j+m)\frac{f_0}{f_s}]\}$$

$$=(\frac{U_r}{2})^2 A_0\{E[\cos(2\pi(2j+m)\ f_0/f_s-\theta_k)]+E[\cos(2\pi m\ f_0/f_s+$$

$$\theta_k)]\}+E[\eta(j)U_r\cos 2\pi(j+m)\ f_0/f_s]$$

$$\approx(\frac{U_r}{2})^2\frac{1}{N}\sum_{j=1}^{N}A_0\cos(2\pi m\ f_0/f_s+\theta_k)+\eta^{(2)}(m)$$

$$=(\frac{U_r}{2})^2 A_0\cos(2\pi m\ f_0/f_s+\theta_k)+\eta^{(2)}(m)$$

where $\eta^{(2)}(m)= E[\eta(j)U_r\cos 2\pi(j+m)\ f_0/f_s]$ and $j=1,2,\cdots N$.

5. REFERENCES

[1] R. W. Sanders, "Digital audio authenticity using the electric network frequency," in Proc. *33rd International Conference: Audio Forensics-Theory and Practice (June 2008)*.

[2] R. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, pp. 84-94, 2009.

[3] C. R. Kriigel, G. Smith, M. Graves, "Audio Analysis," in Proc. *16th International Forensic Science Symposium*, Interpol–Lyon, 5-8 October 2010, pp. 379.

[4] R. Garg, A.L.Varna, M. Wu, "Seeing ENF: natural time stamp for digital video via optical sensing and signal processing," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 23-32.

[5] C. Grigoras, "Digital audio recording analysis the electric network frequency criterion," *International Journal of Speech Language and the Law*, vol. 12, pp. 63-76, 2005.

[6] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Science International*, vol. 167, pp. 136-145, 2007.

[7] A. J. Cooper, "An automated approach to the Electric Network Frequency (ENF) criterion-Theory and practice," *International Journal of Speech Language and the Law*, vol. 16, pp. 193-218, 2009.

[8] M. Huijbregtse and Z. Geradts, "Using the ENF criterion for determining the time of recording of short digital audio recordings," *Computational Forensics*, pp. 116-124, 2009.

[9] D. P. Nicolalde and J. A. Apolinario, "Evaluating digital audio authenticity with spectral distances and ENF phase change," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1417-1420.

[10] D. P. Nicolalde and J. A. Apolinario, L.W.P Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Transactions on Information Forensics and Security*, vol. 5, pp. 534-543, 2010.

[11] Yi-bing Li, XinYue, Xin-yuan Yang, "Estimation of sinusoidal parameters in powerful noise by multi-layer autocorrelation," *Journal of Harbin Engineering University*, vol. 4, p. 027, 2004. (Chinese)

[12] J. Ortega-Garcia, J. Gonzalez-Rodriguez, V. Marrero-Aguiar, "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification," *Speech Communication*, vol. 31, no.2, pp. 255-264, 2000.