Detecting Audio Splicing Forgery Algorithm Based on Local Noise Level Estimation

Xuebo Meng
School of Software Engineering
Xi'an Jiaotong University
Xi'an,China

Chen Li*
School of Software Engineering
Xi'an Jiaotong University
Xi'an,China

Lihua Tian
School of Software Engineering
Xi'an Jiaotong University
Xi'an,China

Abstract—**Splicing including homogenous splicing and heterogeneous splicing is one of the most common ways to tampering digital audio. This paper focuses on the study of heterogeneous splicing. Since different audio is recorded at different background environments, there exist significant differences in background noise level for different recorded audio. In this paper, the level of background noise is used to detect the splicing tampering in audio. However, unlike other methods that use background noise level for tamper detection, we do not detect background noise level after continuous framing. In order to more accurately determine the location of tampering, we first need to determine the length of each syllable in the audio using endpoint detection method. Here, we use the spectral entropy method to determine the length of each syllable. Then the variance of background noise of each syllable is calculated according to the kurtosis value characterized the number of peaks in the density distribution curve at the mean that is close to a constant in the band-pass filtered domain. Finally, by comparing the similarities between the variance of background noise of each syllable, it is judged whether there is an operation of heterogeneous splicing tampering in the audio. Our algorithm can accurately locate the heterogeneous splicing syllables.**

Keywords-splicing; digital audio forensics; background noise

I. INTRODUCTION

With the rapid development of multimedia technology and increased functionality of editing software, it has become easier to tamper digital audio. Electronic proof is an important application of digital multimedia materials in judicial authentication. Therefore, when audio forensics is applied in court forensics, it is necessary to verify the legitimacy, authenticity, and relevance of audio signals.

So far, there are a lot of research results in the field of detecting the manipulation of digital audio. Reference [1] summarizes several techniques for audio forensics and discusses the concepts, models, and advances of these technologies. Overall, audio forensics can be divided into three categories: active forensics technology, audio authenticity forensics technology and audio source identification technology. Audio authenticity forensics technology is divided into two types: authenticity forensics based on electric network frequency characteristics and authenticity forensics based on environmental characteristics. The detection method based on background noise level belongs to the forensic technology based on the environment features in the audio authenticity forensics technology. Reference [3] proposed a digital speech forensics method based on background noise level. It combined geometric transform spectral subtraction and multi-band spectral subtraction to estimate background noise level and then compared the correlation with background noise level in different environments to determine the integrity. Although the method mentioned in this literature can estimate background noise level, it can not locate the tampered position. The algorithm in [4] estimate the various of background noise based on an observed property of audio signals - they tend to have kurtusis that is close to a constant in the band-pass filtered domain and then compare the similarity between the various of background noise to locate the tampered position. However, the location of tampering is not accurate enough. The author of [5] adopts the time recursive average algorithm to estimate the background noise level in the audio signal and then determines whether the audio signal has been tampered by detecting the absence of mutation in the variance of noise by the mutation point detection algorithm.

In this paper, an algorithm for detecting the heterogeneous splicing tampering of audio based on background noise level is proposed. Firstly, the digital audio is detected by the spectral entropy method to detect the position of each syllable in digital audio. After calculating the variance of the background noise of each syllable and calculating the similarity between them, we can locate the specific location of the heterogeneous syllable.

II. THEORY BASIS ABOUT KURTOSIS

A. Kurtosis Constancy

The band-pass filtered domain tends to have a kurtosis that is close to constant, which means that the kurtosis values of the audio signal after passing through different filters are different, but the distribution of the values is more concentrated and can be regarded as a constant. Kurtosis is calculated as follows:

978-1-7281-0120-0/18/$31.00 ©2018 IEEE

$$k = \frac{\mu_4}{(\sigma^2)^2} - 3 \quad (1)$$

k is the audio kurtosis and x is the audio signal sampling values. σ^2 and μ_4 are the second and fourth order central moments of x. Fig.1 illustrates the kurtosis values of the DCT filters of an audio. The audio is convolved with 127 DCT AC filters to produce these kurtosis values. In Fig.1, the red line shows the mean value of the kurtosis after different filters, and the green lines correspond to the mean plus/minus standard deviation respectively.

From Fig. 1, it can be observed that most of the kurtosis values are within a very narrow range between two green lines except for a few values.

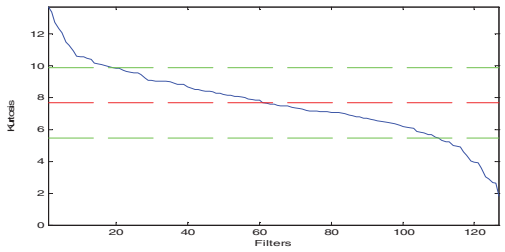

Figure 1. The kurtosis of audio

III. THE PROPOSED METHOD

The proposed detection process mainly includes three parts including endpoint detection, background noise level estimation and similarity comparison. A summary of detect process is shown in Fig. 2.

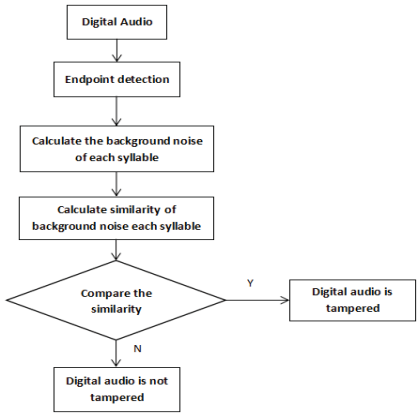

Figure 2. Summary of Detection Process

A. Spectral Entropy Method

The speech signal with the time domain waveform x(n) is framed to obtain the ith frame speech signal $x_i(m)$. Then after Fast Fourier Transform (FFT) that N is the length of FFT, the energy spectrum of the frequency component of the kth spectral line is $Y_i(k)$ .So the normalized spectral probability density function $p_i(k)$ of each frequency component is obtained as follows:

$$p_i(k) = \frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(l)} \quad (2)$$

The short-term spectral entropy H_i of each speech frame is defined as:

$$H_I = -\sum_{K=0}^{N/2} P_I(K) \log P_I(K) \quad (3)$$

For noise, we know that its normalized spectral probability density function is evenly distributed, so its spectral entropy value is large. But for speech signals, since the spectrum has a formant spectrum characteristic, its normalized spectral probability density function is unevenly distributed making the spectral entropy of speech generally lower than the spectral entropy of noise. Therefore, we use this feature to locate speech endpoints from noisy speech.

B. Local Noise Estimation

After endpoint detection, we can locate the position of each syllable. Next, all we have to do is calculating the variance of the background noise of each syllable. The recorded speech y is pure speech x mixed with background noise z. What we need to do is to estimate the variance σ^2 of z from y. First, we convolve the speech signal y with the nth DCT filter from the $1\times N$ basis to generate a response signal y_n. Then according to the [8], we draw the following formula: $\tilde{k}_n = k_n(\frac{\tilde{\sigma}_n^2 - \sigma^2}{\tilde{\sigma}_n^2})^2$. In this equation, $\tilde{k}_n$, k_n and $\tilde{\sigma}_n^2$ respectively represents the kurtosis of y_n , x_n and the variance of y_n. Assume the marginal distribution of band-pass filter responses x_n has super-Gaussian property, i.e. k_n>0, we have $\tilde{k}_n$>0 due to the fact that $\tilde{\sigma}_n^2 > \sigma^2$. Then we take square root on the equation to obtain:

$$\sqrt{\tilde{k}_n} = \sqrt{k_n}(\frac{\tilde{\sigma}_n^2 - \sigma^2}{\tilde{\sigma}_n^2})^2 \quad (4)$$

Assuming k_n are approximately constant across different DCT filters, we can calculate the kurtosis k of pure speech signal x and background noise variance σ^2 by minimizing their squares difference:

$$L(\sqrt{k}, \sigma^2) = \sum_{n=1}^{N} (\sqrt{\tilde{k}_n} - \sqrt{k} + \frac{\sqrt{k}\sigma^2}{\sigma_n^2})^2 \quad (5)$$

According to (5), we get a closed-form optimal solution of (6) and (7):

$$\sqrt{k} = \frac{\left\langle \sqrt{\tilde{k}_n} \right\rangle_n \left\langle \frac{1}{(\tilde{\sigma}_n^{\,2})^2} \right\rangle_n - \left\langle \frac{\sqrt{\tilde{k}_n}}{\tilde{\sigma}_n^{\,2}} \right\rangle_n}{\left\langle \frac{1}{(\tilde{\sigma}_n^{\,2})^2} \right\rangle_n - \left\langle \frac{1}{\tilde{\sigma}_n^{\,2}} \right\rangle_n^2} \tag{6}$$

$$\sigma^2 = \frac{1}{\left\langle \frac{1}{\tilde{\sigma}_n^{\,2}} \right\rangle_n} - \frac{1}{\sqrt{k}} \frac{\left\langle \sqrt{\tilde{k}_n} \right\rangle_n}{\left\langle \frac{1}{\tilde{\sigma}_n^{\,2}} \right\rangle_n} \tag{7}$$

In the above two formula, $\langle \cdots \rangle_n$ means to take the average over different DCT filters.

C. Similarity detection

After calculate the background noise variance of each syllable, we need to calculate the similarity of the background noise variance of any two syllables to determine whether there exists heterogeneous splicing operation. If the similarity is higher than the set threshold, it is considered that there is splicing tampering operation between the two syllables. The formula of absolute difference D between background noise σ_1^2 and σ_2^2 of two syllables as: $D = |\sigma_1^2 - \sigma_2^2|$ and the relative difference $E = \frac{D}{|\sigma_1^2| + |\sigma_2^2|}$. Finally we can obtain the similarity R=1-E as:

$$R = 1 - \frac{|\sigma_1^2 - \sigma_2^2|}{|\sigma_1^2| + |\sigma_2^2|} \tag{8}$$

If the similarity is larger than the set threshold THR, it is determined that the two syllables are recorded in the same environment.

IV. EXPERIMENTS

We prepare 50 audio signals with 44100Hz sampling rate of length around 5 seconds , 50 audio signals with 44100Hz sampling rate of length around 10 seconds and 50 audio signals with 44100Hz sampling rate of length around 30 seconds to do the experiments. Then we add noises of different variances to these 150 audio and randomly spliced some syllables to get 100 tampered audio.

A. Noise Estimation Result

In order to detect the accuracy of noise estimation, we estimate the background noise level of these audio signals added Gaussian white noise convolved with 7 DCT AC filters using method proposed in 3.3. Table 1 summarizes the average performance of this method showing the mean and standard deviation of the estimated noise in SNR. As can be seen from Table 1, the local noise estimation method can accurately estimate the noise in the audio.

B. Endpoint Detection Results

In order to detect the robustness of the method, we select a set of speech and add 10dB white noise as the detection sample. In Fig. 3, (a) is the original speech and (b) is the result of endpoint detection after adding noise. The red solid line is the starting position of the syllable and the blue dotted line is the ending position of the syllables. From Fig.3, it can be seen that the endpoint detection by spectral entropy method can accurately locate the start position and end position of the syllable under difference noise levels.

Then, we use spectral entropy to detect the audio where there exists heterogeneous syllables.

We insert the signal shown in Fig. 4 (a) from 0.6s to 1.7s and 3.6s to 4.95s into the 0.25s and 4.9s of the signal shown in Fig. 4 (b) respectively. Finally, the signal shown in Fig. 4 (c) is formed. The detection results are shown in Fig. 5. In Fig. 5, (a) is the syllable detection results of the audio, (b) is the detection results of the noise variance corresponding to each syllable and (c) is the location results of the heterologous syllable in audio. The results with the method in [4] are shown in the Fig. 6. The red dashed line in Fig. 6 is manually labeled according to the various of background noise.

We insert the signal shown in Fig. 7 (a) from 3.1s to 4.25s into the 4.7s of the signal shown in Fig. 7 (b). Finally, the signal shown in Fig. 7 (c) is formed. In the same way, the audio is detected separately using the method proposed in this paper and the method mentioned in [4]. In Fig. 8, (a) is the syllable detection results of the audio, (b) is the detection results of the noise variance corresponding to each syllable and (c) is the location results of the heterogeneous syllables in audio. The results with the method in [4] are shown in the Fig. 9.

TABLE I. THE AVERAGE PERFORMANCE OF LOCAL NOISE ESTIMATION METHOD

	10dB	15dB	20dB	30dB
5seconds	10.1146	14.9901	20.0093	29.9203
10seconds	10.0209	14.8306	20.0500	29.9541
30seconds	10.0214	14.9899	19.9408	30.0582

The experiment results show that the method in [4] can only roughly determine the position of the splicing part in the audio, but our method can accurately locate the location of the syllable of the splicing part and is not interfered by the meaningless segment.

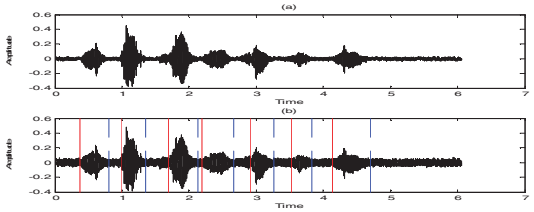

Figure 3. Syllable segmentation (white noise, SNR = 10dB) test results

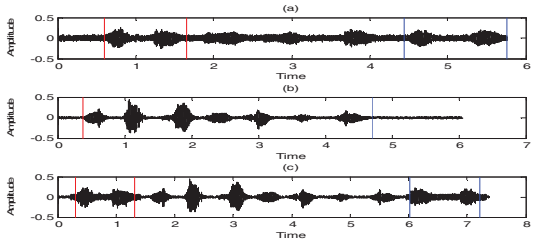

Figure 4. Audio splicing example

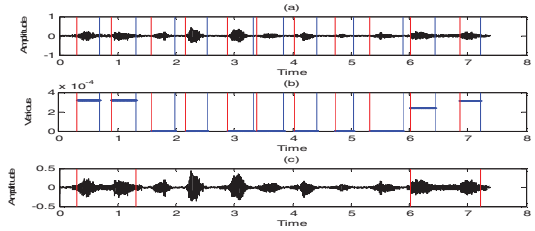

Figure 5. Results of endpoint detection and the variance of the noise

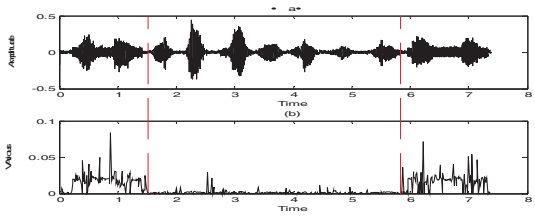

Figure 6. Framing detection results

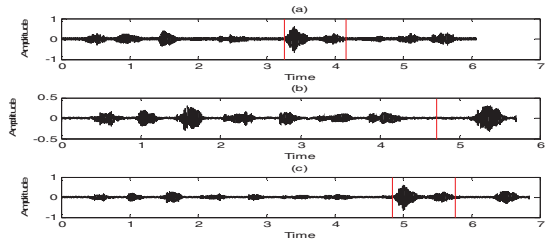

Figure 7. Audio splicing example

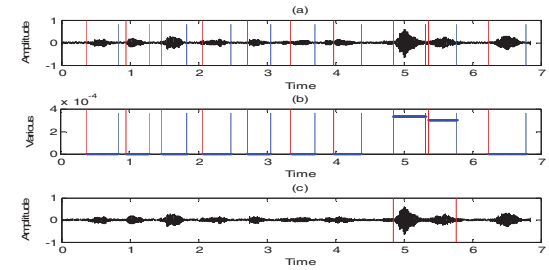

Figure 8. Results of endpoint detection and the variance of the noise

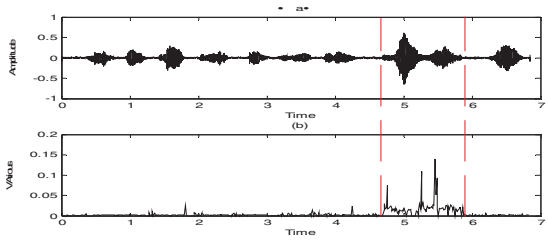

Figure 9. Framing detection results

C. *Tamper detection contrast*

To further illustrate the effect of the method, we use two quantitative measures evaluate the performance of our audio forgery detection method. The sample detection omission (SDO) rate and the sample false positive (SFP) rate are defined as :

$$SDO = \frac{N_1}{N} \tag{9}$$

$$SFP = \frac{N_2}{\widetilde{N}} \tag{10}$$

Here we measured 100 audios. N is the number of authentic original audios. $\widetilde{N}$ is the number of authentic tampered audios. N_1 is the number of audios with the splicing operation is detected as the original audio. N_2 is the number of audios that do not have a splicing operation detected as tampering audio. Here we set the threshold THR to 0.8. The comparison results are provided in Table 2. It can be seen that the accuracy of our method is relatively high. Since the endpoint detection method is adopted in this paper, the location of the tampered syllable can be accurately located and is not interfered by the non-meaningful segment. Therefore, the SDO rate and SFA rate of this method are lower.

V. CONCLUSIONS

In this paper, a heterogeneous splicing detection algorithm is proposed. By calculating the background noise variance of each syllable after endpoint detection, we can locate the position of the heterogeneous syllables. This method can be described as follows. First, we use spectral entropy to detect the position of each syllable. After calculating the variance of the background noise of each syllable and calculating the similarity between them, we can locate the specific location of the heterogeneous syllables. From the experimental results, we can find that our method can more accurately locate the location of heterogeneous syllables compared with method in [4].

TABLE II. COMPARISON RESULTS OF SDA AND SFP

	Our method	Method in [4]
SDO	0.12	0.06
SFA	0.21	0.23

REFERENCES

[1] Yongqiang Bao, Ruiyu Liang, Zhe Cong , et al. Research progress on several key technologies of audio forensics[J]. Data Acquisition and Processing, 2016, 31(2): 252-259.

[2] Anshan Ran, Rangding Wang, Diqun Yan. Mobile phone source identification based on spectrum feature of equipment noise floor [J]. Telecommunication Science, 2017(1):85-94.

[3] Ikram S, Malik H. Digital audio forensics using background noise [C]. Proceedings of the 2010 IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE, 2010:106-110.

[4] Pan X, Zhang X, Lyu S. Detecting splicing in digital audios using local noise level estimation [C]. Proceedings of the 2012 IEEE International Conference on Acoustics. Piscataway, NJ: IEEE, 2010: 1841-1844.

[5] Hua G, Zhang Y, Goh J, et al. Audio authentication by exploring the absolute-error-map of ENF signals [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(5):1003-1016.

[6] Galdo G D. Audio authentication using the kurtosis of ESPRIT based ENF estimates [C]//Proceedings of the 2016 10th IEEE International Conference on Signal Processing and Communication Systems. Piscataway, NJ:IEEE,2016: 1-6

[7] Ling Zou, Qianhua He, Xichao Ruan, et al. Source identification of recording equipment based on equipment noise estimation [J]. Journal of Jilin University (Engineering Science), 2016: 1-8

[8] Huijbregtse M, Geradts Z. Using the ENF criterion for determining the time of recording of Short Digital Audio Recordings [J]. Computational Forensics, 2009, 5718:116-124.

[9] Rui Yang, Zhenhua Qu, Jiwu Huang. Detecting digital audio forgeries by checking frame offsets [C]. International Multimedia Conference, Proceedings of the 10th ACM workshop on Multimedia and security, Oxford, United Kingdom, 2008:21-26.