DETECTING SPLICING IN DIGITAL AUDIOS USING LOCAL NOISE LEVEL ESTIMATION

Xunyu Pan, Xing Zhang and Siwei Lyu

Computer Science Department, University at Albany, SUNY
Albany, NY 12222, USA
{xypan,xz654242,lsw}@cs.albany.edu

ABSTRACT

One common form of tampering in digital audio signals is known as splicing, where sections from one audio is inserted to another audio. In this paper, we propose an effective splicing detection method for audios. Our method achieves this by detecting abnormal differences in the local noise levels in an audio signal. This estimation of local noise levels is based on an observed property of audio signals that they tend to have kurtosis close to a constant in the band-pass filtered domain. We demonstrate the efficacy and robustness of the proposed method using both synthetic and realistic audio splicing forgeries.

Index Terms— Digital Forensics, Audio Splicing, Local Noise Level Estimation

1. INTRODUCTION

Digital audios have become ubiquitous with the popularity of the internet and portable digital devices such as personal music players and smartphones. In the meanwhile, rapid developments of low-cost and sophisticate editing software make the modification of audio file much easier for untrained users. There have been several recent cases of audio forgery drawing the public's attention, including the alleged tampering of the recorded audio of actor Mel Gibbson [7], and the controversy over the authenticity of the audio files claimed to be the voices of Osama Bin Laden [8]. The increasing number of forged audios calls for more effective tools for the authentication and forgery detection for digital audios.

In this paper, we describe a new method that can be applied to detect a common form of tampering in digital audio signals known as *splicing*, where sections from one audio are inserted into another audio. Our method achieves this by detecting abnormal differences in the local noise levels in an audio signal. The estimation of local noise levels is based on an observed property of audio signals – they tend to have kurtosis close to a constant in the band-pass filtered domain. The variance of noise in the audio signal is estimated by minimizing an objective function that has a closed-form optimal solution. We examine the noise level inconsistency within the audio file, which are used to detect the location and length of suspicious audio clips. We also report the robustness and effectiveness of our method using both synthetic and realistic audio forgeries with splicing tampering.

2. PREVIOUS WORK

Digital watermarking may be used to protect the authenticity of audio [3, 11]. However, to apply digital watermarking, it is necessary to have particular hardware/software support that most non-professional digital audios recording devices lack.

Recent years, several active forensic detection methods for audio signals have been developed [4, 5, 6, 12]. For instance, acoustic device, e.g. microphones, are identified [4, 5] by extracting background features of audio stream. Similar forensic tool based on the amount of sound reverberation, which uniquely decides the shape and composition of a room where the audio signal was recorded, is proposed in [6]. In another work [12], the digital tampering in MP3 audio data is identified by checking the inconsistency of frame offsets.

However, most of these methods assume some knowledge of the recording device or the specific file format. On the other hand, we may obtain more general forgery detection methods using common statistical properties of digital audios independent of specific recording devices or file formats.

In [10], local noise levels are estimated by computing the second and fourth moments at each local signal block. But the method assumes that the kurtosis values of the original signal are known, which is hard to satisfy in practice. Our proposed work is most closely related with the work in [13], where the optimal values for the kurtosis of the original clean signal and the variance of the noise are sought simultaneously by minimizing an objective function, assuming scale invariance of signal kurtosis. In contrast, our method has an efficient implementation based on a closed-form solution, and can be extended to estimate local noise levels.

3. METHOD

3.1. Kurtosis Constancy

The audio kurtosis κ, which represents the peakedness of the distribution of the signal sampling values x, is defined as $\kappa =$

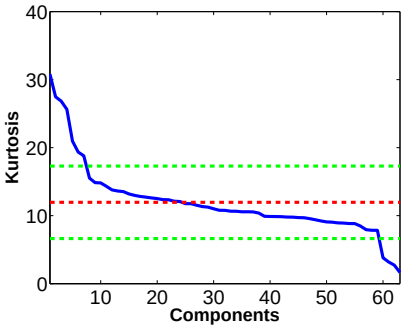

Fig. 1. *Kurtosis profiles - kurtosis values (blue line) as a function of the component number, or frequency, of the 63 DCT AC filters for an audio from the TIMIT database. The red line corresponds to the mean of the kurtosis values, and the green lines correspond to the mean plus/minus standard deviation respectively.*

$\frac{\mu_4}{(\sigma^2)^2}$-3, where σ^2 and μ_4 are the second and fourth order central moments of x. Fig.1 illustrates the kurtosis values of the DCT responses of an audio selected from the TIMIT dataset [1], a high quality clean corpus of phonemically and lexically transcribed speech of American English speakers. More specifically, the audio is convolved with 63 DCT AC filters to produce the responses, where the kurtosis values of these response signals are computed and sorted. It can be observed from Fig.1 that, except for a few outliers, most of the kurtosis values fall into a narrow range around their mean value[1]. Based on this observation, we describe an efficient method to estimate the variance of noise added to clean audio signals.

3.2. Global Noise Level Estimation

As the basis of our audio splicing detection method, we first introduce a global noise estimation algorithm for audio signals based on the near constancy property of their kurtosis values in DCT domain. Let us denote a clean audio signal $\mathbf{x}$, and $\mathbf{y} = \mathbf{x} + \mathbf{z}$ as the result of $\mathbf{x}$ contaminated by a additive white Gaussian noise (AWGN) $\mathbf{z}$ of unknown variance σ^2. Our goal is to estimate σ^2 from $\mathbf{y}$. To this end, we produce the response signal y_k by the convolution of y with the kth filter from the $1 \times N$ DCT basis. We further denote κ_k, $\tilde{\kappa}_k$, and $\tilde{\sigma}_k^2$ as the kurtosis of x_k and y_k, and the variance of y_k, respectively. The kurtosis of x_k and y_k, and the variance of y_k and σ^2 are related as [9]: $\tilde{\kappa}_k = \kappa_k \left(\frac{\tilde{\sigma}_k^2 - \sigma^2}{\tilde{\sigma}_k^2} \right)^2$. Assume the marginal distributions of band-pass filter responses, x_k has super-Gaussian property, i.e. $\kappa_k > 0$, we have $\tilde{\kappa}_k > 0$ due to the fact that $\tilde{\sigma}_k^2 > \sigma^2$. We can further take square root on the equation to obtain:

$$\sqrt{\tilde{\kappa}_k} = \sqrt{\kappa_k} \left(\frac{\tilde{\sigma}_k^2 - \sigma^2}{\tilde{\sigma}_k^2} \right). \tag{1}$$

Assuming κ_k are approximately constant across different DCT bands, we can estimate the kurtosis of the audio signal κ and its variance σ^2 by by minimizing their squared difference:

$$L\left(\sqrt{\kappa}, \sigma^2 \right) = \sum_{k=1}^{N^2} \left(\sqrt{\tilde{\kappa}_k} - \sqrt{\kappa} + \frac{\sqrt{\kappa}\sigma^2}{\tilde{\sigma}_k^2} \right)^2. \tag{2}$$

Eq.(2) has a closed-form optimal solution, as:

$$\sqrt{\kappa} = \frac{\left\langle \sqrt{\tilde{\kappa}_k} \right\rangle_k \left\langle \frac{1}{(\tilde{\sigma}_k^2)^2} \right\rangle_k - \left\langle \frac{\sqrt{\tilde{\kappa}_k}}{\tilde{\sigma}_k^2} \right\rangle_k \left\langle \frac{1}{\tilde{\sigma}_k^2} \right\rangle_k}{\left\langle \frac{1}{(\tilde{\sigma}_k^2)^2} \right\rangle_k - \left\langle \frac{1}{\tilde{\sigma}_k^2} \right\rangle_k^2} \tag{3}$$

$$\sigma^2 = \frac{1}{\left\langle \frac{1}{\tilde{\sigma}_k^2} \right\rangle_k} - \frac{1}{\sqrt{\kappa}} \frac{\left\langle \sqrt{\tilde{\kappa}_k} \right\rangle_k}{\left\langle \frac{1}{\tilde{\sigma}_k^2} \right\rangle_k}, \tag{4}$$

where we use $\langle \cdot \rangle_k$ as a shorthand notation for the average over different band-pass filters.

It should be noted that even though we assume AWGN, this is not as restricted as it seems, as very non-Gaussian independent noise in the time domain will mix in to be Gaussian noise in the filter domain due to the central limit theorem and noise independence.

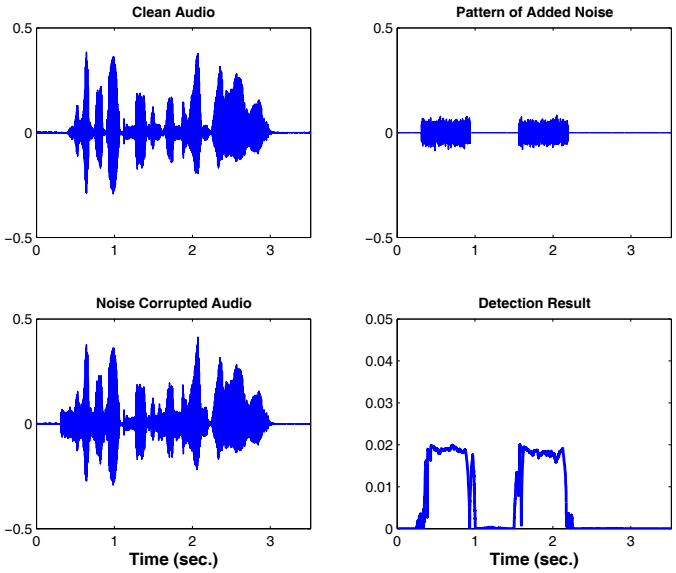

Fig. 2. *Illustration of local noise level estimation on a noise corrupted audio signal.*

3.3. Local Noise Level Estimation

The global noise level estimation method can be further extended for the estimation of locally varying noise levels in audio signals. Instead of straightforward implementation using sliding estimation windows, we propose a more efficient local noise estimation algorithm based on dynamic programming. In particular, we term the *integral vector* of an audio signal x, denoted as $\mathcal{A}(\mathbf{x})$, is an audio signal of the same size as $\mathbf{x}$, but its value at index i is the sum of all sampling values of $\mathbf{x}$ in the range of $[1, i]$. The integral vector can be efficiently

[1] Similar observations can be made for other types of band-pass linear filters.

constructed, and it can be used to compute the sum in any sub-interval $[i, i+I]$ with subtraction operation as $\mathcal{A}(\mathbf{x})_{i+I} - \mathcal{A}(\mathbf{x})_i$. Specifically, we can efficiently compute the k^{th} order spatial statistics for a time interval $\Omega = [i, i+I]$, as:

$$\mu_k(\mathbf{x}_\Omega) = \frac{1}{\Omega}\left[\mathcal{A}(\underbrace{\mathbf{x} \circ \cdots \circ \mathbf{x}}_{k \text{ times}})_{i+I} - \mathcal{A}(\underbrace{\mathbf{x} \circ \cdots \circ \mathbf{x}}_{k \text{ times}})_i\right], \quad (5)$$

where $\circ$ represents point-wise multiplication. Furthermore, the local variance and kurtosis of a surrounding time interval Ω of each sampling location can be computed as:

$$\kappa(\mathbf{x}_\Omega) = \frac{\mu_4(\mathbf{x}_\Omega) - 4\mu_3(\mathbf{x}_\Omega)\mu_1(\mathbf{x}_\Omega) + 6\mu_2(\mathbf{x}_\Omega)\mu_1(\mathbf{x}_\Omega)^2 - 3\mu_1(\mathbf{x}_\Omega)^4}{\left(\sigma^2(\mathbf{x}_\Omega)\right)^2} - 3 \quad (6)$$

$$\sigma^2(\mathbf{x}_\Omega) = \mu_2(\mathbf{x}_\Omega) - (\mu_1(\mathbf{x}_\Omega))^2. \quad (7)$$

We can substitute the results of Eqs.(6) and (7) into Eqs.(3) and Eqs.(4) to find the local noise level in the audio signal. Shown in Fig.2 is the pipeline of the proposed local noise level estimation method, which can be further employed for audio forgery detection.

4. EXPERIMENTS

In this section, we first test the performance of the proposed global audio noise estimation algorithm presented in Section 3.2. We collect a set of 452 high quality audio signals with 16KHz sampling rate of length around 3 seconds from the TIMIT dataset [1]. The noise corrupted signals are generated by adding AGWN of different variances to these original signals. To accommodate the different dynamic ranges of these signals, we measure noise level with the signal-to-noise ratio (SNR). We estimate the global noise for each corrupted audio signal with 7 DCT AC filters using the algorithm described in Section 3.2. As a comparison, we also implement a 1D noise variance estimation method for audio based on the iterative numerical estimation algorithm presented in [13] with the same 7 DCT AC filters. Table 1 summarizes the average performance of both methods, showing the mean and standard deviation (in parenthesis) of the estimated noise levels in SNR for the same set of corrupted audio signals. The average running time on the TIMIT audio dataset using our method is only 0.03 seconds, which is much shorter than the 0.24 seconds average running time for the audio noise estimation method converted from the algorithm described in [13].

	10 dB	15 dB	20 dB	30 dB
Work in [13]	10.45(0.68)	15.20(0.83)	19.71(1.21)	27.83(4.28)
Our method	10.04 (0.13)	15.01 (0.24)	20.00(0.41)	30.73(3.53)

Table 1. *The average performance of our noise estimation method, with comparisons of the performance of the work in [13] on various levels of additive white noises (signal-to-noise ratio: SNR).*

We next evaluate the performance of the local audio noise level estimation algorithm described in Section 3.3. A set of

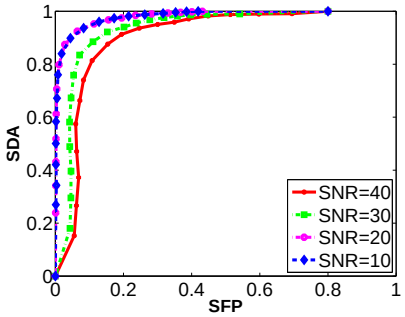

Fig. 3. *ROC curves for the average SDA and SFP rates on 100 noise corrupted audio signals with different SNRs.*

noise corrupted audio signals are generated using randomly selected audio signals in the TIMIT dataset. We corrupt each audio signal with white Gaussian noise at a chosen time interval of 20% of the length of the signal, with various noise levels. For each noise level, we generate 100 tampered audio signals. We then use two quantitative measures to evaluate the performance of our audio forgery detection method. Denote Ω as the region of audio samples in the true noise tampered time interval, and $\tilde{\Omega}$ as the region of audio samples in the detected region. We define the *sample detection accuracy* (SDA) rate as the fraction of audio samples in noise corrupted region that are correctly identified, i.e., SDA $= \frac{|\tilde{\Omega} \cap \Omega|}{|\Omega|}$ and the *sample false positive* (SFP) rate as the fraction of audio samples in uncorrupted region that are detected falsely as from noise corrupted time interval, i.e., SFP $= \frac{|\tilde{\Omega} - \Omega|}{|\bar{\Omega}|}$. To reduce the effect of random samples, each pair of SDA/SFP rates is computed as the averages over all 100 corrupted audios at each noise level. The resulting ROC curves for various added noise levels are shown in Fig.3. Note that the overall detection accuracies for lower SNRs are better, because higher noise levels tend to be easier to detect. In addition, we also test our method on quantization noise due to rounding in digital signal processing, and the estimation algorithm still holds.

As a further demonstration of the effectiveness of the proposed method, we create realistic audio forgeries using audio editing software GoldWave from GoldWave Inc.[2]. More specifically, we select two source audio clips in WAV format with sampling rate 16KHz and data rate 256Kbps from the TIMIT dataset. We also select an original sound track in MP3 format with sampling rate 22KHz and data rate 24Kbps for the popular episode *The Marine Biologist* of American television sitcom *Seinfeld*. For fair comparison, we first compress the TIMIT WAV file to MP3 file with the same sampling and data rates as those of the *Seinfeld* MP3 sound track. We next perform two experiments with insertion or substitution of chosen word segments from the *Seinfeld* episode into the two compressed TIMIT audio clips. In the first experiment as shown in Fig.4, we crop a phrase chosen from the *Seinfeld* episode and insert it at the beginning of one TIMIT audio clip. During the manipulation process, we carefully choose the tampering section so that the resulting sentence is still

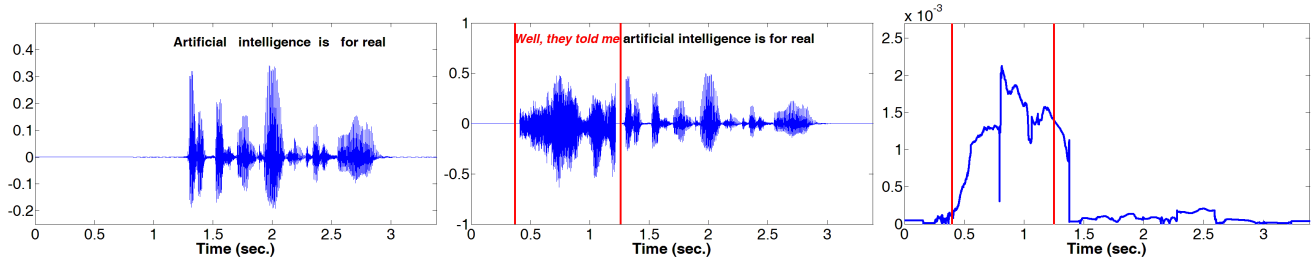

Fig. 4. *Example of insertion tampering, where the phrase "*Well, they told me*" cropped from an episode of* Seinfeld *is inserted into a source TIMIT audio signal (left). In both forged audio signal (middle) and detection result (right), the locations of the inserted audio segment are marked in red.*

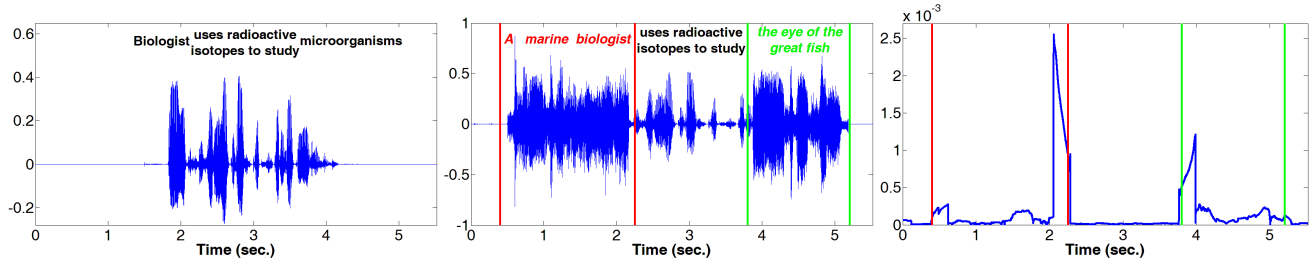

Fig. 5. *Example of substitution tampering, where the words "*biologist*" and "*microorganisms*" in a source TIMIT audio signal (left) are replaced by "*A marine biologist*" and "*the eye of the great fish*" cropped respectively from an episode of* Seinfeld*. In both forged audio signal (middle) and detection result (right), the locations of the two substituted audio segments are marked in red and green respectively.*

meaningful. We also tune both the volume and speed of the splicing audio components using GoldWave[2] to make the forged audio signal hears more realistic. In the second experiment as shown in Fig.5, we substitute two words in another TIMIT audio signal for two word segments cropped from the episode of *Seinfeld*. The detection results demonstrate that the individual splicing segments in the forged audio signal exhibit significant noise level differences, which provides strong evidence of audio tampering.

5. DISCUSSION

In this work, we describe a fast and blind local noise level estimation method, which can be further employed to detect digital audio forgeries. Our method does not require specific knowledge of the recording devices or the file format, and is sufficiently efficient for forgery inspection on a large repository of audio files. We are currently in the process of developing a reliable splicing detection method based on the estimated local noise levels.

6. REFERENCES

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM, 1993. National Institute of Standards and Technology, NISTIR 4930.

[2] GoldWave Inc., www.goldwave.com. *GoldWave v5.58*.

[3] Kirovski and Malvar. Spread-spectrum watermarking of audio signals. *IEEE Transactions on Signal Processing*, 51(4):1020–1033, 2003.

[4] Christian Krätzer, Andrea Oermann, Jana Dittmann, and Andreas Lang. Digital audio forensics: a first practical evaluation on microphone and environment classification. In *ACM MM&Sec*, Dallas, TX, 2007.

[5] Christian Krätzer, Kun Qian, Maik Schott, and Jana Dittmann. A context model for microphone forensics and its application in evaluations. In *Proceedings of SPIE Media Watermarking, Security, and Forensics III*, San Francisco, CA, 2011.

[6] Hafiz Malik and Hany Farid. Audio forensics from acoustic reverberation. In *ICASSP*, Dallas, TX, 2010.

[7] ABC News. Did someone mess with mel gibson's audio recordings? `http://abcnews.go.com/Entertainment/mel-gibsons-rants-messed/story?id=11169736`, July 2010.

[8] BBC News. Bin laden tape 'not genuine'. `http://news.bbc.co.uk/2/hi/middle_east/2526309.stm`, November 2002.

[9] D. R. Pauluzzi and N. C. Beaulieu. A comparison of snr estimation techniques for the AGWN channel. *IEEE Transactions on Communications*, 48(10):1681–1691, 2000.

[10] A.C. Popescu and H. Farid. Statistical tools for digital forensics. In *6th International Workshop on Information Hiding*, Toronto, Canada, 2004.

[11] Shijun Xiang and Jiwu Huang. Histogram-based audio watermarking against time-scale modification and cropping attacks. *IEEE Transactions on Multimedia*, 9(7):1357–1372, 2007.

[12] Rui Yang, Zhenhua Qu, and Jiwu Huang. Detecting digital audio forgeries by checking frame offsets. In *ACM MM&Sec*, Oxford, UK, 2008.

[13] Daniel Zoran and Yair Weiss. Scale invariance and noise in nature image. In *ICCV 2009*, Kyoto, Japan, 2009.