COPY-MOVE DETECTION OF AUDIO RECORDING WITH PITCH SIMILARITY

Qi Yan[1], Rui Yang[2], Jiwu Huang[3]*

[1]School of Information Science and Technology, Sun Yat-sen University, China
[2]School of Information Management, Sun Yat-sen University, China
[3] College of Information Engineering (Shenzhen Key Laboratory of Media Security), Shenzhen University
Email: yangr23@mail.sysu.edu.cn

ABSTRACT

The widespread availability of audio editing software has made it very easy to create forgeries without perceptual trace. Copy-move is one of popular audio forgeries. It is very important to identify audio recording with duplicated segments. However, copy-move detection in digital audio with sample by sample comparison is invalid due to post-processing after forgeries. In this paper we present a method based on pitch similarity to detect copy-move forgeries. We use a robust pitch tracking method to extract the pitch of every syllable and calculate the similarities of these pitch sequences. Then we can use the similarities to detect copy-move forgeries of digital audio recording. Experimental result shows that our method is feasible and efficient.

Index Terms—Audio forgeries, Audio forensics, Copy-Move detection, Pitch similarity

1. INTRODUCTION

Nowadays digital audio recordings have played an important role in digital evidence. Unfortunately, the ease manipulation of digital audio recordings has made its authenticity often in doubt. Imagining the following situation, someone used audio editing software to copy some segments of audio recording and pasted these segments into other positions of the same audio. Then the original meaning of the audio may be changed. Because the magnitude and frequency of the same words will not change radically, these copy-move words will be imperceptible. After some post-processing, such as adding noise, filtering, recompressing on the tempered speech, it will be very difficult to find the duplicated segments. Thus copy-move detection of digital audio recording is an urgent issue of audio forensics.

In recent years, the authenticity of digital audio has got many researchers' attention. Brian[1] used frequency spectrum analysis to detect MP3 bit rate quality. Grigoras[2] and Garg[3] used the electric network frequency (ENF) to check the integrity of digital audio. Farid[4] used bispectral analysis to detect audio forgery. There are also some existing works about copy-move detection of image [5-6]. However, few works about copy-move detection of digital audio recording are reported. Xiao[7] presented a method using the similarity between audio segments to detect duplicated segments. The technology applied in audio fingerprinting[8] may be useful for copy-move detection in digital audio. However, audio fingerprinting methods are used for music retrieval, the granularity of audio clips is often larger than 5 seconds.

In this paper, we propose a new method to detect the duplicated segments in digital audio. We use a robust method to extract pitch sequences of every syllable and calculate the similarities of pitch sequences. The duplicated segments are identified as the ones with high similarities.

The rest of this paper is organized as follow. In Section 2, we briefly introduce the pitch tracking method, YAAPT[9]. In Section 3, we introduce our detection process including pitch extraction and comparison. In Section 4, experimental results show that our method is feasible and efficient. Finally, in Section 5 we give the conclusion and discussion about this paper and future work.

2. PITCH AND PITCH TRACKING METHOD

Since audio recordings used for evidence in courts are often long, detecting copy-move in audio sample by sample is very time consuming. After some post-processing, it's also difficult to detect copy-move forgeries sample by sample. Hence, the solution for copy-move detection of audio recording should fulfill two requirements: low computation and robustness against post-processing. After extensive experiments and theoretical analysis, we found the pitch sequence satisfy these requirements. Thus, in this paper, we use pitch sequence to identify duplicated segments.

2.1.Basic Knowledge about Pitch

Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale[10]. Pitch is usually quantified as a frequency and refers to fundamental

This work was support by National Science and Technology Pillar Program (2012BAK16B06), NSFC (61202497),the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and Shenzhen R&D Program (GJHZ20140418191518323).

frequency. Pitch sequences extracted from different syllables are often quite different. Even a person says identical words twie, the pitch sequences extracted from the syllables are very different. An example is shown in Fig. 1. There are three "Two" in the sentence. The 1st and 3rd "Two" are original, while the 2nd "Two" is a copy of the 3rd. As We can see, the 2nd and the 3rd "Two" have similar pitch sequences, while the pitch sequence of the 1st "Two" is quite different.

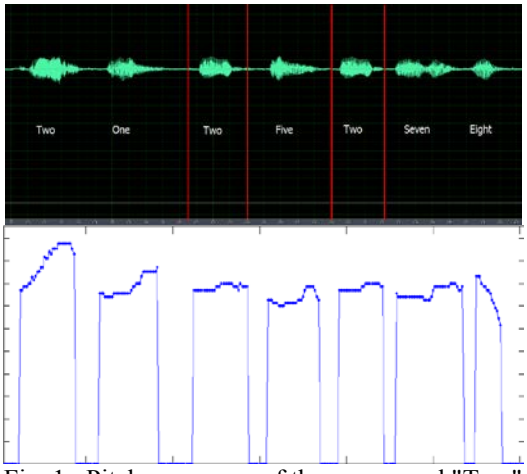

Fig. 1 Pitch sequences of the same word "Two"

2.2.Pitch Tracking Method

Pitch tracking is a classic problem of audio processing. There are many traditional pitch tracking methods, such as Auto-correlation Function (ACF), Average Magnitude Difference Function (AMDF) and Cepstrum. But traditional pitch tracking methods may have some disadvantages, so we use a robust pitch tracking method, YAAPT[9].

YYAPT is mainly based on the normalized cross correlation function[11] (NCCF) and spectrum analysis. This algorithm mainly includes four steps:

(1). Audio signal preprocessing. Nonlinear processing of the signal is used to create multiple versions of the signal.

(2). Spectrum analyzing to extract pitch. We use the spectral correlation (SHC) to extract the pitch and use the normalized low frequency energy ratio (NLFER) to distinguish voiced frames and unvoiced frames.

(3). NCCF is applied to extract pitch candidates from both original signal and nonlinearly processed signal.

(4). Dynamic programming is introduced to get the final pitch from the pitch candidates we got in step 2 and step 3.

3. PROPOSED METHOD

3.1.Overview of Detection Process

The proposed detection process mainly includes two parts, pitch extraction and comparison of pitch sequences. A summary of detect process is shown in Fig. 2.

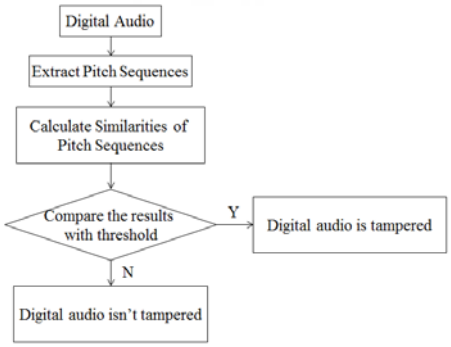

Fig. 2 Summary of Detection Process

3.2.Pitch Extraction

In Sec 2, this paper has briefly introduced YAAPT. Here, we mainly describe the implementation of pitch extraction. This method includes 4 steps:

(1). We use nonlinear processing to get multiple versions of signal. In this paper, We use squared value of the signal as the nonlinear processing of the signal.

(2). SHC is applied to extract pitch. The spectral harmonics correlation is defined as follows[9]:

$$SHC(t,f)=\sum_{f'=-WL/2}^{WL/2}\prod_{r=1}^{N_H+1}S(t,rf+f') \qquad (1)$$

Where $S(t,f)$ means the magnitude spectrum for frame t at frequency f, N_H represents the numbers of harmonics, WL represents spectral window length. $S(t,f)$ is normalized to [0,1], and will has high amplitude at integer multiples of pitch.

NLFER is used as an aid for pitch extracting. The NLFER is defined as follows:

$$NLFER(t)=\frac{\sum_{f=2\times F_{0\_\min}}^{F_{0\_\max}}S(t,f)}{\frac{1}{T}\sum_{t=1}^{T}\sum_{f=2\times F_{0\_\min}}^{F_{0\_\max}}S(t,f)} \qquad (2)$$

Where T is the total number of frames. $S(t,f)$ is the spectrum for frame t and frequency f. NLFER is used to distinguish voiced frames and unvoiced frames. NLFER will have high amplitude for voiced frames and low amplitude for unvoiced frames.

(3). NCCF is used to select pitch candidates from original and nonlinearly processed signals. The NCCF is defined as follows:

$$NCCF(k)=\frac{1}{\sqrt{e_0 e_k}}\sum_{n=0}^{N-K\_\max}s(n)s(n+k) \qquad (3)$$

Where N is frame length of the signal $s(n)$, $e_0=\sum_{n=0}^{N-K\_\max}s^2(n)$ and $e_k=\sum_{n=0}^{k+N-K\_\max}s^2(n)$. K_min and K_max represent the lag value used to accommodate pitch tracking range. NCCF is normalized to [-1,1], and will get max value at integer multiples of pitch.

(4). In step 2 and step 3, we get a series of pitch candidates, then we use dynamic programming to choose the result of pitch.

3.3.Comparison of Pitch Sequences

We have extracted pitch sequences for every syllable. In this paper we use Pearson Correlation Coefficient (PCCs) and the average difference (AD) of pitch sequences as similarity measure.

(1). Pearson Correlation Coefficient

In this paper we choose Pearson Correlation Coefficient to calculate the similarities at the trend level of every pitch sequences. PCCs is defined as follows:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X^2)}{N})(\sum Y^2 - \frac{(\sum Y^2)}{N})}} \qquad (4)$$

Where X and Y are corresponding values of two sequences, N is the length of sequence. PCCs is normalized to [-1,1], and two sequences are similar at the trend level, when their PCCs is close to 1. In Fig. 3, we calculate PCCs of every syllable in a tampered audio. The 48th to 51st segments are copied from 20th to 23rd segments. It is obvious that the pitch sequences of these segments have high PCCs.

	46	47	48	49	50	51	52
16	0.3874	-0.3050	-0.7647	0.8400	0.5836	-0.3814	-0.3012
17	-0.8078	0.0126	0.0656	0.1573	-0.2873	0.1865	0.2005
18	0.6095	0.0394	-0.8815	0.9339	0.4347	-0.3375	-0.3366
19	0.4708	-0.0254	-0.8207	0.8977	0.4501	-0.3410	-0.3740
20	-0.8536	-0.5901	1	-0.9002	-0.2507	-0.3146	0.3822
21	0.7443	0.7666	-0.8939	0.9829	0.4194	0.5308	-0.1738
22	0.8195	0.0838	-0.2904	0.4619	0.9808	-0.1284	-0.1521
23	-0.7328	0.2437	-0.2110	0.5110	-0.1715	0.7969	0.0759
24	-0.4461	-0.3357	0.2587	-0.2876	-0.1656	-0.0570	0.3414

Fig. 3 PCCs of Tampered Audio

(2). Average Difference

We use the Average Difference (AD) to compare two pitch sequences at the value level. AD is defined as follows:

$$AD = \frac{1}{N_{\min}} \sum_{n=1}^{N_{\min}} \left| x(n) - y(n) \right| \qquad (5)$$

Where $N_{\min}$ is the length of short sequences, $x(n)$ and $y(n)$ are two pitch sequences. Values of tow sequences are closer, the result of AD is smaller. In Fig. 4, we calculate AD of pitch sequences in a tampered audio. The 48th to 52nd segments are copied from 20th to 24th segments. We can find that the AD of duplicated segments is much less than that of normal segments. We can choose an appropriate threshold to distinguish two pitch sequences at value level.

	45	46	47	48	49	50	51	52	53	54
16	5.7162	26.2774	15.2900	20.7004	56.9065	17.2516	41.6163	68.2548	59.4436	43.1092
17	47.5584	46.0380	45.4635	30.9047	57.4943	76.6619	22.1700	26.2384	32.3220	28.0658
18	54.9269	82.0656	61.5835	40.7538	34.0461	48.2340	50.1409	14.3071	13.0671	46.8229
19	10.7456	34.3575	8.1271	9.3467	55.0897	34.3864	42.9512	57.0891	52.7579	51.8218
20	17.0061	41.0663	37.8321	0	77.0954	63.0864	36.2017	42.0503	50.7466	27.3225
21	51.1325	53.0114	78.2717	75.4536	2.2337	59.9859	78.2214	33.9503	24.7397	25.7318
22	15.6511	11.7182	41.7807	63.9672	56.9914	2.9307	52.3995	69.7166	53.8333	59.0129
23	41.0272	19.1848	40.1604	39.6688	75.4244	54.4357	1.7592	51.0317	55.9894	47.8457
24	64.8813	91.1531	65.5015	40.0274	37.5860	68.7597	50.0206	2.7724	4.9222	43.0847
25	59.7148	85.4018	72.3666	50.4621	21.8425	58.8165	56.8866	4.2103	3.8618	48.8040
26	5.2453	21.1897	5.5174	5.5174	21.8137	24.3796	26.3187	52.3922	50.4477	24.3634
27	12.9829	15.2358	21.0684	53.6984	91.8994	28.5855	31.2835	79.3517	79.3005	87.6781
28	24.4527	17.5738	29.4000	39.7937	72.0923	59.2054	17.7119	62.5851	54.8397	74.8711

Fig. 4 AD of Tampered Audio

4. EXPERIMENTAL RESULTS

4.1.Audio Database and Generation of Tampered Audio

We prepare 1000 tampered audios with duplicated segments. Time of every audio is longer than 30 seconds. Each duplicated segment is 0.6s to 2s. Then we make post-processing, such as adding noise, filtering, recompressing and get 3000 post-processed tempered audios. The audio database used in this experiment consists of these audios.

Each audio used in this experiment is mono, the sampling rate is 8kHz and the format is WAV file.

4.2.Selection of Threshold

In order to verify whether pitch can distinguish duplicated segments and normal segments, we make some experiments. We calculate the PCCs and AD of the pitch sequences of the segments that one person repeat at different time. The result is shown in TABLE I and TABLE II. It's obvious that even the same person repeats the same word at different time, the pitch sequences are quite different. Thus, we can use pitch to identify duplicated segments.

TABLE I. PCCs of repeated segments

PCCs	Over 0.90	Over 0.85	Over 0.80	Over 0.75	Over 0.70	Over 0.65
Percentage	0.017%	0.103%	0.349%	1.712%	3.203%	6.073%

TABLE II. AD of repeated segments

AD	Under 2.0	Under 2.5	Under 3.0	Under 3.5	Under 4.0	Under 4.5
Percentage	0.000%	0.000%	0.013%	0.037%	0.062%	0.102%

In order to choose appropriate threshold, we calculate PCCs and AD of pitch sequences of audio in audio database. The statistics are shown in TABLE III and TABLE IV. TABLE III is the statistical result of PCCs. It's obvious that duplicated segments have higher PCCs meanwhile normal segments have lower PCCs. TABLE IV is the statistical result of AD. It's obvious that we can use AD to distinguish duplicated segments from normal segments at the value level. According to the statistical result and experimental result, we choose 0.78 as threshold of PCCs and choose 3.6 as threshold of AD.

TABLE III. Statistical Result of PCCs

PCCs	Over 0.90	Over 0.85	Over 0.80	Over 0.75	Over 0.70	Over 0.65
Duplicated Segments	88.915%	93.812%	99.314%	99.731%	99.902%	99.994%
Normal Segments	2.874%	8.117%	14.772%	21.739%	28.903%	39.942%

TABLE IV. Statistical Result of AD

AD	Under 2.0	Under 2.5	Under 3.0	Under 3.5	Under 4.0	Under 4.5
Duplicated Segments	84.127%	93.079%	97.592%	99.398%	99.891%	99.997%
Normal Segments	0.435%	0.616%	0.793%	1.025%	1.287%	1.593%

4.3.Detection Example

In this section, we use a detection example to explain detection process. The audio to be detected is a speech recording. Length is 32.607 seconds and the sampling rate is 8kHz. There is one duplicated segments in this audio. Fig. 5 shows this tampered segment in time domain. The segments between red lines are duplicated segments.

Fig. 5 Example of copy-move detection of audio

We extract pitch sequences of this audio and finally get 101 pitch sequences. Then we calculate PCCs and AD of every two pitch sequences. Fig. 6 is PCCs of the tampered segment and Fig. 7 is AD of the tampered segment. It's obvious that the 48th to 51st sequences have high PCCs and low AD with the 20th to 23rd sequences. This means that the 48th to 51st sequences and the 20th to 23rd sequences are duplicated.

	45	46	47	48	49	50	51	52
16	0.6384	0.3874	-0.3050	-0.7647	0.8400	0.5836	-0.3814	-0.3012
17	-0.7408	-0.8078	0.0126	0.0656	0.1573	-0.2873	0.1865	0.2005
18	0.6027	0.6095	0.0394	-0.8815	0.9339	0.4347	-0.3375	-0.3366
19	0.1785	0.4708	-0.0254	-0.8207	0.8977	0.4501	-0.3410	-0.3740
20	-0.8204	-0.8536	-0.5901	1	-0.9002	-0.2507	-0.3146	0.3822
21	0.8412	0.7443	0.7666	-0.8939	0.9829	0.4194	0.5308	-0.1738
22	0.6531	0.8195	0.0838	-0.2904	0.4619	0.9808	-0.1284	-0.1521
23	-0.4666	-0.7328	0.2437	-0.2110	0.5110	-0.1715	0.7969	0.0759
24	-0.3832	-0.4461	-0.3357	0.2587	-0.2876	-0.1656	-0.0570	0.3414

Fig. 6 PCCs of the First Tampered Segment

	45	46	47	48	49	50	51	52
16	5.7162	26.2774	15.2900	20.7004	56.9065	17.2516	41.6163	68.2548
17	47.5584	46.0380	45.4635	30.9047	57.4943	76.6619	22.1700	26.2384
18	54.9269	82.0656	61.5835	40.7538	34.0461	48.2340	50.1409	14.3071
19	10.7456	34.3575	8.1271	9.3467	55.0897	34.3864	42.9512	57.0891
20	17.0061	41.0663	37.8321	0	77.9554	63.9864	36.2017	42.0503
21	51.1325	53.0114	78.2717	75.4536	2.2337	59.9859	78.2214	33.9503
22	15.6511	11.7182	41.7807	63.9672	56.9914	2.9307	52.3995	69.7166
23	41.0272	19.1848	40.1604	39.6688	75.4244	54.4357	1.7592	51.0317
24	64.8813	91.1531	65.5015	40.0274	37.5860	68.7597	50.0206	2.7724

Fig. 7 AD of the First Tampered Segment

4.4.Detection Result in Audio Database

This section mainly introduces the detection results in audio database. We use the true positive rate (TPR) and the false positive rate (FPR) as the standards to evaluate the detection result. The result is as TABLE VII shows.

TABLE V. Detection Result

Types of audio	Original	Adding noise	Filtering	Recompressing
TPR	99.962%	98.514%	99.398%	99.748%
FPR	0.498%	0.987%	0.763%	0.627%

The detection result in audio database proves our method to detect duplicated segments is feasible and efficient. By choosing suitable threshold of PCCs and AD, we can get high TPR and low FPR.

In a practical application, the audio recording used for evidence is usually very long. Searching duplicated segments sample by sample costs much time. For example, we detect the audio in section 4.3. Searching duplicated segments sample by sample cost 46.237s but our method only costs 6.051s. This means that our method can detect duplicated segments quickly.

In order to prove our method is feasible in practical application, we make statistics of the average running time of this method. TABLE VI is the result. The result indicates the detection method presented by this paper is also feasible in practical application.

TABLE VI. Average Running Time

Time of audio	30s	1min	1h	2h
Average Running Time	6.462s	17.134s	0.674h	1.619h

4.5.Comparison with Other Detection Method

There are only a few of researches about detecting duplicated segments in audio are achieved. Here we choose a classical audio fingerprinting system[8] to detect duplicated segments in audio and make comparison with our method. We use TPR, FPR and average running time to compare these two methods. TABLE VII is the comparison of TPR and FPR. TABLE VIII is the comparison of average running time. From these comparisons, it's obvious that our method has better TPR and FPR and costs less time compared with using audio fingerprinting.

TABLE VII. Comparison of Two Methods

Method	Using pitch sequences		Using audio fingerprinting	
	TPR	FPR	TPR	FPR
Original	99.962%	0.498%	95.147%	0.741%
Adding noise	98.514%	0.987%	90.193%	1.379%
Filtering	99.398%	0.763%	92.716%	1.074%
Recompressing	99.748%	0.627%	93.058%	0.951%

TABLE VIII. Comparison of Average Running Time

Method / Time	Using pitch sequences	Using audio fingerprinting
30s	6.462s	20.431s
1min	17.134s	57.932s
1h	0.674h	1.412h
2h	1.619h	3.718h

5. CONCLUSION

In this paper, a copy-move detection method of digital audio recording is proposed. Based on comparison of pitch sequences, we can detect duplicated segments in the audio and locate the positions of duplicated segments. This method can be described as follows. First, we use a robust pitch tracking method to exact pitch sequences of every syllable. After calculating PCCs and AD of every sequence and comparing PCCs and AD with the threshold, we can detect duplicated segments and locate the positions of duplicated segments. From the experimental results, we can find that our method can effectively detect duplicated segments in audio and located the positions of duplicated segments. Compared with audio fingerprinting, our method has higher TPR and lower FPR and also costs less time.

To improve accuracy and efficiency of detection is our future work. Furthermore, we will try to extract more information of the audio as characteristics to detect the duplicated segments in the next stage.

REFERENCES

[1] B. D'Alessandro，Y. Shi，MP3 Bit Rate Quality Detection through Frequency Spectrum Analysis，International Multimedia Conference,Proceedings of the 11th ACM workshop on Multimedia and security Princeton，2009.

[2] C.Grigoras，Digital Audio Recording Analysis: The Electric Network Frequency (ENF) Criterion，The International Journal of Speech Language and the Law，12（1), pp:63-76，2005.

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

[4] H. Farid, Detecting Digital Forgeries Using Bispectral Analysis，MIT AI Memo AIM-1657, MIT，1999.

[5] J. Fridrich, B. Soukal, and A. Lukáš. "Detection of copy-move forgery in digital images." in Proceedings of Digital Forensic Research Workshop, 2003.

[6] A. Irene. “A sift-based forensic method for copy–move attack detection and transformation recovery.” Information Forensics and Security, IEEE Transactions on 6(3), pp: 1099-1110, 2011.

[7] J. Xiao J, Y. Jia, E. Fu. Audio authenticity: Duplicated audio segment detection in waveform audio file. Journal of Shanghai Jiaotong University (Science), 2014, 19: 392-397.

[8] J. Haitsma, T. Kalker. A highly robust audio fingerprinting system. ISMIR 2002, pp:107-115.

[9] A. Stephen，H. Hu，A spectral/temporal method for robust fundamental frequency tracking，The Journal of the Acoustical Society of America, vol 123, pp:4559-4571，2008.

[10] A. Klapuri and M. Davy.Signal processing methods for music transcription. Springer, ISBN 978-0-387-30667-4,2006

[11] D. Talkin，Arobust algorithm for pitch tracking (RAPT)，Speech coding and synthesis, pp:495- 518,2005