ISBN 978-93-84422-22-6
Proceedings of 2015 International Conference on Image Processing, Production and Computer Science (ICIPCS'2015)
Istanbul (Turkey), June 3-4, 2015 pp. 75-82

Robust Tampered Detection Method for Digital Audio using Gabor Filterbank

Fajri Kurniawan[1], Mohammed S. Khalil[1] and Hafiz Malik[2]

[1]Center of Excellence Information Assurance, King Saud University, Riyadh, Saudi Arabia
[2]Electrical and Computer Engineering Department, University of Michigan – Dearborn, Dearborn, MI 48128, USA

fkurniawan.c@ksu.edu.sa, sayimkhalil@gmail.com, hafiz@umich.edu

***Abstract*:** *Nowadays, audio editing tools can easily utilized to alter any digital audio signal. The original recorded conversation can be modified by inserting fake statement in order to twisted the context. Most of such tampering are difficult to identify by relying only on human hearing. Hence, a robust tool is required to help detecting tampered audio if present. In forensic community, it is known that digital traces exists on each audio signal due to characteristics of the acquisition device. Detecting the acquisition device information can be helpful for forensic practitioner to evaluate consistency of the recording. In this study, three features are investigate to classify the microphone models while take into consideration the issue of identical model. Those features are analyzed and compared in the experiments. The result indicated that Gabor filterbank feature outperformed than others. Thus, the Gabor feature has great potential to localize forgery that present on digital audio recording.*

Keywords: *audio tampered detection, fingerprint, gabor filterbank, mfcc, plp, microphone forensic*

1. Introduction

Digital multimedia are widely used in many website or online media to deliver information more attractively. It also used for many purposes such as evidence in court, entertainment industry, health and medical and other public interest. Such digital content can be formed from speech/conversation recording, camera footage and photograph. The raw content must contain authentic information as it capture the real-time situation. However, after post-processing the information might altered or twisted when someone intentionally tamper the raw content. Hence, an forgery detection can be useful to investigate altered digital content.

Currently, intensive research in forgery detection is mainly focus in digital image and video content. Meanwhile, it is still lack of study in audio forgery detection. This paper would study on audio forgery technique. The audio forgery techniques are very useful to combat digital forgery problems such as copyright issue, blurring court evidence, alter recording of figure or politician statement for black campaign, authenticate correctness of financial recording and other sensitive issue.

As an example, the corruption issue that raise in developing country, where it involving large amount of money, the corruptors may talk over the phone when dealing with some party. The investigator can reveal the corruption once it happen through the taped conversation. However, other party can try to obscure the evidence somehow by altering the taped conversation with another conversation, which recorded using identical phone. Hence, investigator have to proof it in the court that the recording was genuine using some forgery detection technique. In another case, an record company sometimes distribute low quality version of an album for

promotional purposes. The copyright pirates would utilized that low quality album and somehow make over it into fake high quality album and sell it to make profit. The customer and record company both would suffer losses. Hence, the audio forgery technique is greatly demanded these days.

Related works in forgery detection for audio signal can be found in [1-16, 24]. Farid [10] proposed bispectral analysis to authenticate forgery in human speech. Using bispectral analysis, he assumed un-natural higher order statistical correlation would occurs in forged signal. However, this technique only suitable when the major tampering happen on the signal. Yang et al [6, 7, 11] studied audio forgery based on frame offsets but it works only for MPEG audio format. Maher [4] described forensics tools on audio enhancement and interpretation. Nicolalde & Apolinario [8] utilized electrical network frequency (ENF) to authenticate digital audio recording. Yang [24] argued that recent forgery detection method would failed once post-processing applied on forged audio. Yang proposed white noise detection based on sign change rate approach. Unfortunately, that method very limited to detect additive noise only. Meanwhile, other researchers works on audio forensic by exploring microphone identification [12, 13], reflected sounds [14] and audio coding analysis [15,16], but none of them study tampered audio that involved identical microphone model.

As discuss in literature, more works are still required in audio forgery detection because many issues are remain open. Talking about sophisticated forgery, they would attempt to hide audible trace due to forgery using some techniques such as applying some noises or filters at post-processing. As instance, the forger could adding white noise after altered the audio signal to blur all traces. Such task can be done with less effort using audio editing software that available in market, for example Audacity. However, it is obvious detecting various of noises and filters at a time will time consuming and superfluous. Hence, this study proposed a novel approach to detect the forgery by considering microphone's fingerprint. Audio signal that have unstable fingerprint can be considered as tampered audio signal. This paper investigated three features named Gabor filterbank, MFCC and PLP to identify the microphone model. The method is tested tampered audio that recorded using 5 microphone models, where each one at least have two to three identical models. The comparison is carried out to reveal the best features for microphone's fingerprint.

The paper is structured as follows: In Section 2.1, the method that utilized to extract the microphone's fingerprint is described. Afterward, the dataset is presented in section 2.2. Finally, experimental result and conclusion are shown in Section 3 and Section 4, respectively.

2. Proposed Method

2.1. Microphone's Fingerprint Extraction

In general, microphone's fingerprint is derived from digital traces that exists on digital audio signal due to several factors [19]. In this work three features are considered as microphone's fingerprint, which are Gabor filterbank, mel-frequency ceptral coefficient (MFCC) and perceptual linear prediction (PLP). Accordingly, there is three different features that represent as a fingerprint of each microphone. Initial assumption is each microphone has unique fingerprint such that twelve microphones presented by twelve fingerprints. The fingerprint is simply extracted from audio signal with fixed window size. The window size is set to 0.025 seconds in the experiment. Figure 1 depicted the general overview of proposed method to obtain the microphone's fingperprint.

MFCC and PLP can be considered as baseline on the comparison because those features have been widely used and proven robust in speech recognition [22, 23]. In general, PLP and MFCC shared several analogous steps. First, both features applied hamming window and DFT on input signal. Then, a set of filterbank are employed to generate the power spectrum. Obviously, the main difference between PLP and MFCC are on filterbank that utilized and how processing the power spectrum to produce 13 features set. For the detailed MFCC and PLP features can be read in reference [22, 23].

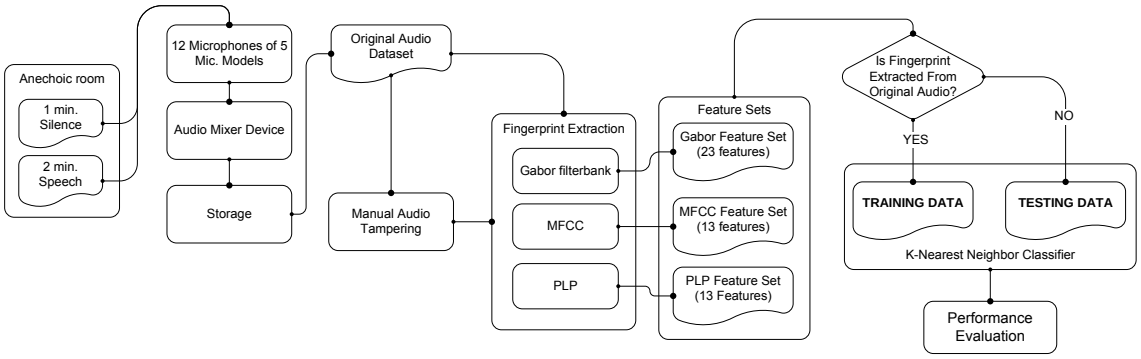

Fig. 1: Diagram of Proposed Feature Extraction for Microphone's Fingerprint

The Gabor filterbank would briefly described in this paper since this feature still new in audio signal processing. Schadler et al. [20] proposed Gabor filterbank Auditory Spectrum (GFAS) features for speech recognition. The Gabor function is considered as proven able to represent speech signal in compact spectro-temporal structure, allow constant overlap, compress the output excess and has flexible parameters. Schadler et al reported that their features are robust for speech recognition. Hence, this study attempt to exploit Gabor filterbank feature on tampered audio detection. The flowchart of Gabor filterbank feature extraction is depicted in figure 2.

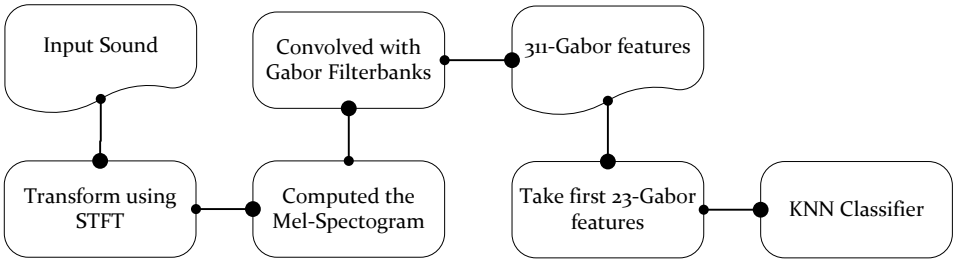

Fig. 2: Diagram of Proposed Feature Extraction

1. Perform Short Time Fourier Transform (STFT) equation 1 on fixed window part of input signal to transform it into power spectrum, as denoted below:

$$X_m(w) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-jwn} = DTFT_w\left(x.Shift_{mR}(w)\right) \quad (1)$$

Where, $x(n)$ and $w(n)$ are input signal at time n and length of M window function (e.g. Hamming) respectively. $X_m(w)$ and *R* are DTFT function of segmented data that centralized with time *mR* and hop size in the middle of consecutive DTFTs, respectively.

2. The Mel-spectrogram is generated based on ETSI standard [21]. It contain twenty three frequency channels that divided into equal distance called are lower, central and higher frequency. The Mel-spectogram is calculated as follow:

$$\boldsymbol{H}(f) = \begin{cases} 0, f \leq fl \text{ and } f \geq fh \\ \dfrac{(f - fl)}{(fc - fl)}, fl \leq f \leq fc \\ \dfrac{(f_h - f)}{(fh - fc)}, fc \leq f \leq fh \end{cases} \quad (2)$$

3. The GFAS feature then extracted from the generated Mel-spectogram by applying 59 2D-Gabor filter that has shaped from two parameter spectral and temporal values [20]. The filter is denoted as $g(.)$ and computed below:

$$h_b(x) = \begin{cases} 0.5 - 0.5\, cos\left(\frac{2\pi x}{b}\right), & if\ -\frac{b}{2} < x < \frac{b}{2} \\ 0, & else \end{cases} \tag{3}$$

$$s_w(x) = exp(iwx) \tag{4}$$

$$g(k_0, n_0, w_k, w_n, k, n, v_k, v_n, \emptyset) = s_{wk}(k - k_0) s_{w_n}(n - n_0).\, h_{\frac{v_k}{2w_k}}(k - k_0) h_{\frac{v_k}{2w_n}}(n - n_0).\, e^{i\emptyset} \tag{5}$$

The parameter including: k is channel and n is time-frame, it centralized at k_0 and n_0, spectral and temporal modulation frequency denoted as w_k and w_n, respectively, then number of semi-cycles determined using v_k and v_n. Schadller et al [20] defined various parameters to produce 59 filters that named GFAS. The 311 features are generated by convolving those filterbank with input signal in mel-spaces.

4. In this study we reduced the 311-features into only first 23-features based on observation that first few filters are actually the basis for another filters. Hence, the first few features hold basic features of the input signal. This reduction also decrease the classifier burden thus improve the computation cost.

2.2. Dataset

Fourteen audio recording are collected prior to generate tampered audio files. As mention before in Section 1, this study taking into account identical microphone model exists on tampered audio. In this regards, five different models consist of Shure SM-58, Electro Voice RE-20, Coles 4038 Ribbon, Sennheiser MD421II Dynamic Cardioid Microphone and AKG C 451 B Condenser are utilized, where each of them at least has two identical model. Table I presented short information for each microphones including model name, number of microphone and their naming convention. Twelve microphones of five microphone models are arranged at anechoic room. Such room is reverberation free and free from surrounding noises. This room is chosen to ensure the recorded audio produce clear signal.

The data collection using those twelve microphones are recorded simultaneously by organizing the microphone using microphone stand such that it well-organized. The recording session has three minutes recording, where at first minute the microphone recorded silence sound. Then, remaining two minutes a person is reading a paragraph that contain some sentences repeatedly until 2 minute exceed. There is fixed 30 cm distance between the person's lips and the microphones. Table II depicted audio sample description. Finally, the tampered audio files are generated by replacing "destination file" at particular time with audio signal taken from "source file" as described in Table III.

TABLE I: Microphone Features And Specifications

Shure SM-58 (Mic_1: 3 units)	**Electro-Voice RE-20 (Mic_2: 2 units)**	**Coles 4038 Ribbon Microphone (Mic_3: 2 units)**	**Sennheiser MD421II Dynamic Cardioid Microphone (Mic_4: 3 units)**	**AKG C 451 B Condenser Microphone (Mic_2: 2 units)**
				
SHU_0058	ELE_0020	COL_4038	SEN_0421	AKG_0451

TABLE II: The Audio Sample Description

Description	**Value**
Format	Wave
Audio Format	PCM
Codec ID	1
Bit rate	705.6 Kbps
Channel(s)	1 channel
Sampling rate	44.1 KHz
Bit depth	16 bits
File size	~16.9 MB
Overall bit rate mode	Constant
Bit rate mode	Constant
Format settings, Endianness	Little
Format settings, Sign	Signed

TABLE III: Tampered Description

No.	**File Name**	**Short Name**	**Source File**	**Destination File**	**Start Time**	**End Time**
1	T_AKG_0451_m1_m2	T_AKG	AKG_0451_m2	AKG_0451_m1	0:01:12	0:01:19
2	T_COL_4038_m2_m1	T_COL	COL_4038_m1	COL_4038_m2	0:02:26	0:02:45
3	T_ELE_0020_m1_m2	T_ELE	ELE_0020_m2	ELE_0020_m1	0:01:21	0:01:38
4	T_SEN_0421_m1_m2	T_SEN_A	SEN_0421_m2	SEN_0421_m1	0:02:36	0:02:53
5	T_SEN_0421_m3_m1	T_SEN_B	SEN_0421_m1	SEN_0421_m3	0:01:03	0:01:19
6	T_SHU_0058_m1_m3	T_SHU_A	SHU_0058_m3	SHU_0058_m1	0:02:09	0:02:24
7	T_SHU_0058_m2_m3	T_SHU_B	SHU_0058_m3	SHU_0058_m2	0:01:38	0:01:56

3. Experimental Result

A preliminary study is performed on tampered database to study tampered detection based on fingerprint features. The experiment is conducted by grouping same models into two or three classes depend on corresponding number of microphone of the model. Intra-class problem is the main concern in this experiment. Hence, the classifier will be burdened with less number of classes such that can reveal the robustness of each features under identical model. As mention on Section 2, all three features called Gabor filterbank, MFCC and PLP will be compared. Those feature extraction methods are applied on both original and tampered dataset. Afterward, train-data and test-data are fairly constructed through 10-fold cross validation. The K-NN classifier is utilized to classify the microphone model definitely after it trained with the train-dataset. The classifier recognition rates on seven tampered audio are presented in Figure 3 and Table IV.

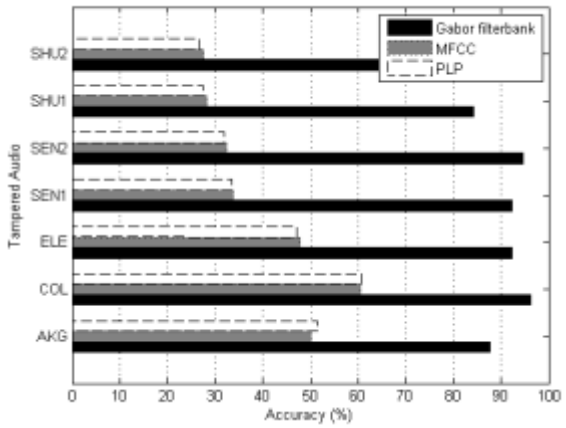

Fig. 3: Accuracy of K-NN Classifier Identify the Tampered Audio.

TABLE IV: Accuracy (in %) of K-NN Classifier

Description	T_AKG	T_COL	T_ELE	T_SEN_A	T_SEN_B	T_SHU_A	T_SHU_B
gabor: 23f_2970d	**87.73**	**96.21**	**92.38**	**92.39**	**94.59**	**84.3**	**85.78**
mfcc: 13f_2980d	50.13	60.55	47.81	33.86	32.38	28.2	27.46
plp: 13f_2980d	51.48	60.75	47.19	33.49	31.86	27.64	26.58

As depicted in figure 1, the extracted features length are as follow: reduced-GABOR has 23 features [20], MFCC has 13 features [22] and PLP has 13 features [23]. As shown in above depicted tables and figure, it is clearly shows that 23 features of reduced-Gabor filterbank obtained highest accuracy followed by 13 features of MFCC and last is PLP feature. From table IV, the highest rates using reduced-Gabor feature can obtained 96.21% on tampered audio of COL_4038 model. This result proven that microphone's fingerprint based on reduced-Gabor has strong discriminator to recognize the microphone model even it come from identical model. Accuracy of reduced-Gabor on other models are also promising with at least 84% correct rates. On the other hand, MFCC and PLP shows very poor and not robust as a fingerprint. According to Table IV, MFCC and PLP can achieved maximum correct rates not greater than 61%. In addition, both features even give very low accuracies less than 34% correct rates for model SEN_0421 and SHU_0058.

4. Conclusion & Future Works

Tampering digital audio signal is ease these days using any simple to sophisticated audio editing tools. Hence, there is also great demand to detect such forgery or to verify the originality of some digital audio signal. This paper compared three features that exploited as microphone's fingerprint to identify tampered audio. According to experimental result, the tampered audio can be identified with high correct rates using Gabor filterbank features. On the other hand, the common features used in audio signal named MFCC and PLP shown not robust when used as a microphone's fingerprint. The result shows that inconsistency in audio recording can be detected based on the digital traces even though it taken from identical model. The Gabor filterbank feature outperform with accuracy of 96.21% at speech recording. Meanwhile, other features MFCC and PLP are almost produce same accuracy not more than 60.75%.

This study can be extended to study blind forgery detection. In such case, no prior knowledge is required to locate the tampered region. Meanwhile, the proposed method is still required preliminary training data to detect the forgery. In addition, more study can be carried out on various places covered indoor and outdoor environment. It is interesting to know more how echo, reverberant or any noise can affect the digital traces.

5. Acknowledgements

This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-INF2634-02).

6. References

[1] F. Rumsey, ‘‘Forensic Audio Analysis,’’ J. Audio Eng. Soc., vol. 56, no. 3, 2008, pp. 211-217.

[2] B.E. Koenig and D.S. Lacey, ‘‘Forensic Authentication of Digital Audio Recordings,’’ J. Audio Eng. Soc., vol. 57, no. 9, 2009, pp. 662-695.

[3] J. Tibbitts and L. Yibin, ‘‘Forensic Applications of Signal Processing,’’ IEEE Signal Processing Magazine, vol. 26, no. 2, 2009, pp. 104-111.
http://dx.doi.org/10.1109/MSP.2008.931099

[4] R.C. Maher, ‘‘Overview of Audio Forensics,’’ Intelligent Multimedia Analysis for Security Applications, 1st ed., H.T. Sencar et al., eds., Springer, 2010, pp. 127-144.
http://dx.doi.org/10.1007/978-3-642-11756-5_6

[5] E.B. Brixen, ‘‘Techniques for the Authentication of Digital Audio Recordings,’’ Proc. AES 122nd Convention, Audio Eng. Soc., 2007, paper 7014.

[6] R. Yang, Z. Qu, and J. Huang, ‘‘Detecting Digital Audio Forgeries by Checking Frame Offsets,’’ Proc. 10th ACM Workshop Multimedia and Security, ACM Press, 2008, pp. 2-26.
http://dx.doi.org/10.1145/1411328.1411334

[7] R. Yang, Q.Y. Shi, and J. Huang, ‘‘Detecting Double Compression of Audio Signal,’’ Proc. SPIE 7541, SPIE Press, 2010; doi:10.1117/12.838695.
http://dx.doi.org/10.1117/12.838695

[8] D. Nicolalde and J. Apolinario, ‘‘Evaluating Digital Audio Authenticity with Spectral Distances and ENF Phase Change,’’ Proc. IEEE Int’l Conf. Acoustics, Speech and Signal Processing, IEEE CS Press, 2009, pp. 1417-1420.
http://dx.doi.org/10.1109/icassp.2009.4959859

[9] C. Grigoras, “Digital audio recording analysis: The electric network frequency (enf) criterion,” The International Journal of Speech Language and the Law, vol. 2, no. 1, pp. 63–76, 2005.
http://dx.doi.org/10.1558/sll.2005.12.1.63

[10] H. Farid, “Detecting digital forgeries using bispectral analysis,” MIT AI Memo AIM-1657, MIT, 1999.

[11] R. Yang, Z. Qu, and J. Huang, “Exposing mp3 audio forgeries using frame offsets,” ACM Transactions on Multimedia Computing, Communications and Application, vol. 8, no. S2, pp. 35:1–35:20, 2012.

[12] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, “Digital audio forensics: A first practical evaluation on microphone and environment classification,” in Proc. of the Workshop on Multimedia and security, Dallas, Texas, USA, 2007, pp. 63–74.
http://dx.doi.org/10.1145/1288869.1288879

[13] R. Buchholz, C. Kraetzer, and J. Dittmann, “Microphone classification using fourier coefficients,” in Proc. of the International Workshop on Information Hiding, Darmstadt, Germany, June 2009.
http://dx.doi.org/10.1007/978-3-642-04431-1_17

[14] H. Malik and H. Farid, “Audio forensics from acoustic reverberation,” in Proc. of the International Conference on Acoustics Speech and Signal Processing, march 2010, pp. 1710–1713.
http://dx.doi.org/10.1109/icassp.2010.5495479

[15] S. Hicsonmez, E. Uzun, and H.T. Sencar, “Methods for identifying traces of compression in audio,” in Proc. of the1st International Conference on Communications, Signal Processing, and their Applications, Sharjah, 2013, pp. 1–6.
http://dx.doi.org/10.1109/iccspa.2013.6487284

[16] R. Yang, Y. Q. Shi, and J. Huang, “Detecting double compression of audio signal,” in Proc. of SPIE 7541, Media Forensics and Security II, 2010.

[17] H. Yuan, “Blind forensics of median filtering in digital images,” IEEE Transactions on Information Forensics and Security, vol. 6, no. 4, pp. 1335–1345, 2011.
http://dx.doi.org/10.1109/TIFS.2011.2161761

[18] G. Cao, Y. Zhao, and R. Ni, “Detection of image sharpening based on histogram aberration and ringing artifacts,” in Proceedings of IEEE International Conference on Multimedia and Expo, 2009, pp. 1026–1029.

[19] Zhao, H., Malik, H., Audio Recording Location Identification using Acoustic Environment Signature. IEEE Transactions on Information Forensics and Security, vol. 8(11), pp. 1746 – 1759, Nov. 2013.
http://dx.doi.org/10.1109/TIFS.2013.2278843

[20] Schadler, M.R., Meyer, B.T. and Kollmeier, B., Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition, Journal of the Acoustical Society of America, 2012
http://dx.doi.org/10.1121/1.3699200

[21] ETSI Standard 201 108 v1.1.3 2003

[22] Logan, B., "Mel Frequency Cepstral Coefficients for Music Modeling." ISMIR. 2000.

[23] Hermansky, H., Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America 87, 1738 (1990)
http://dx.doi.org/10.1121/1.399423

[24] Rui Yang, "Additive noise detection and its application to audio forensics," Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA) , vol., no., pp.1,5, 9-12 Dec. 2014.
http://dx.doi.org/10.1109/apsipa.2014.7041688