Digital Signal Processing 60 (2017) 63–74

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Exposing speech tampering via spectral phase analysis

Xiaodan Lin [a,b], Xiangui Kang [a,*]

[a] Guangdong Key Lab of Information Security, School of Data and Computer Science, Sun Yat-Sen University, 510006, Guangzhou, China
[b] School of Information Science and Engineering, Huaqiao University, 361021, Xiamen, China

ARTICLE INFO

Article history:
Available online 9 August 2016

Keywords:
Spectral phase reconstruction
Tampering localization
Short-time Fourier Transform
Higher order statistics
Spectral phase correlation

ABSTRACT

Audio recordings serve as important evidence in law enforcement context. The most crucial problem in practical scenarios is to determine whether the audio recording is an authentic one or not. For this task, blind audio tampering detection is typically performed based on electric network frequency (ENF) artifacts. In case there is a high level of noise, ENF analysis would become invalid. In this paper, we present a novel approach to detect and locate tampering in uncompressed audio tracks by analyzing the spectral phase across the Short Time Fourier Transform (*STFT*) sub-bands. Spectral phase reconstruction is employed to counteract the impact of noise. Also, a new feature based on higher order statistics of the spectral phase residual and the spectral baseband phase correlation between two adjacent voiced segments is proposed to allow for an automated authentication. Experimental results show that a significant increase in detection accuracy can be achieved compared to the conventional ENF-based method when the audio recording is exposed to a high level of noise. We also testify that the proposed method remains robust under various noisy conditions.

1. Introduction

In the past decade, multimedia forensics had emerged as a hot topic in the field of information security. Earlier, more efforts were dedicated to image forensics since popular image processing software like Adobe Photoshop can be easily grasped by an amateur in image processing. Later, forensic issues got extended to audios and videos as well, due to the availability of editing tools, *e.g.* Audio Audition, Adobe Premiere for those intended to forge the audio or video, whether with or without malicious content manipulation.

Unlike image forgeries, it is much easier to forge an audio by cutting, insertion, substitution or splicing without being noticed even by well-trained ears. The main reason for this may lie in the fact that silence (unvoiced segment) appears constantly in speech signals, thus facilitating local tampering. For local image forgeries, post-processing such as rotating, resizing, sharpening, blurring are always required, aiming to make the image look more natural. However, these post-processings also leave more telltale signs to forensic investigators, leading to technological evolutions of forensics and anti-forensics [1,2]. Several investigations stem from those post-processing footprints such as [3], where an effective median-filtering detector was presented. In [4], contrast enhancement was detected by checking the pixel value histogram.

Despite that it seems much easier to tamper with an audio, it is never an easy task to identify and localize a digital audio recording that has undergone manipulations. Nowadays, audio forensics has covered topics like double compression [5,6], fake MP3 bitrate detection [7], compression history identification [8,9], etc. These forgeries are often global and content-preserving. Hence, all the audio segments exhibit similar variations and the resulting statistical features are available to machine learning techniques. It should be noted that the methods provided in [5–9] are designed specifically to reveal the MP3 compression history, relying on the unique features introduced in the process of MP3 encoding, such as the traces left in the quantized modified discrete cosine transform coefficients. However, there is still an urgent need for the detection of audio content tampering, which seems more appealing to law enforcement agencies. The major challenge for audio content authentication comes from its local characteristics, as audio tampering is performed within targeted fragments of the audio. For example, a keyword or a syllable is cropped from the acoustic signal, leading to misconceptions or ambiguity of the content. For the purpose of local tampering detection, most of the-state-of-art methods utilize the electric network frequency (ENF) signal [10–12]. The random fluctuation of ENF signals across time and different geological locations endows audio signals with unique ENF patterns, and hence can be taken as a type of environmental signatures. All these methods require recovering the ENF signals accurately. However, if the audio is corrupted by high levels of noise, accurate extraction of ENF signals becomes difficult and the performance deteriorates

* Corresponding author.
E-mail address: isskxg@sysu.edu.cn (X. Kang).

http://dx.doi.org/10.1016/j.dsp.2016.07.015
1051-2004/

rapidly, needless to mention the cases without explicit ENF signals, *e.g.*, mobile devices don't directly carry ENF signals. In addition to the footprint left by the power grids, acoustic reverberation which varies depending on the shape and the composition of a room can also be regarded as environmental signatures. Related works can be found in [13–15]. Other than the environmental traces, device fingerprints can also expose audio manipulation, *e.g.*, microphone identification [16].

In the literature, the issue of audio tampering localization is mostly addressed with the aid of authentication codes or watermarks [17–19]. Different from these methods, blind tamper detection achieves the same task without the need of using extrinsic information. Existing works include authenticating waveform audio recordings by detecting ENF phase discontinuity [20,21]. The basic idea is that local audio tampering will violate the steady phase variation of ENF signals, as the normal ENF fluctuation is expected to exhibit a pseudo-periodic pattern. For MP3 format files, the integrity can be identified by checking frame offsets [22]. In [23], Pan et al. proposed a method to localize audio splicing by estimating local noise level. Some extensions to [20] can be seen in [24,25]. In [24], by comparing maximum cross correlation between the extracted ENF signal and the reference signal blockwise, better localization accuracy was yielded. As to [25], the authors proved the viability of using superior harmonic of the ENF signal to evaluate audio authenticity. Another audio forgery localization approach using the singularity with wavelet was given in [26]. However, the noisy condition was not taken into consideration in [26] and false negative error increased if the number of forge operation was small. The authors in [27] combined the technique of microphone classification with the ENF analysis to detect tampered audios. Another recent work operating on ENF abnormality was reported in [28], where the authors employed a data-driven threshold-based strategy to deal with the anomalous variations of the ENF signal. In particular, the difference between the extracted ENF and its median-filtered version highlighted the ENF abnormality, which was then captured by a Two-Pass Split-Window. The same authors of [28] explored ENF patterns to implement the task of audio edit detection [29]. Both the methods were demonstrated to outperform its counterparts in terms of detection accuracy for audio recordings with favorable noise conditions. However, the profile under noisy conditions remained unsatisfactory.

In this work, we focus on the tough problem encountered by most forensic examiners, that is, the audio isn't tampered with in a global sense. In particular, the forgers usually concentrate on the content rather than the signal itself. To tackle this type of fraud, we present a detection method based on spectral phase analysis. We further demonstrate that even under a high level of external noise, the recovered spectral phase can still be applied to audio forensics. The rationale for locating audio tampering will also be revealed.

The rest of this paper is organized as follows. In Section 2, we analyze why the ENF-based methods failed under noisy conditions, followed by a revelation of the rationale behind the STFT analysis for phase reconstruction in Section 3. In Section 4, an approach for localizing tampered speech via spectral phase analysis is presented. Evaluations of the proposed method on both clean and noisy audio recordings are given in Section 5. Finally, conclusions are drawn in Section 6.

2. ENF analysis under noisy conditions

First we will have a review on the conventional tampering detectors based on the ENF signals. Without loss of generality, we assume that the ENF signal is coupled with the speech in the process of recording. Thus, the speech signal can be formulated by

$$y(n) = s(n) + f(n), \tag{1}$$

where $s(n)$ denotes the genuine speech, and the ENF signal is denoted by $f(n)$. It is known that ENF signals are nominally with frequencies fluctuating around 50 Hz or 60 Hz [11]. Hence, $y(n)$ is the recorded signal in which the ENF signal is incorporated. All of the detection algorithms based on ENF have to firstly detach $f(n)$ from $y(n)$. A common practice is applying a narrow band-pass filter centered at the nominal frequency to the questioned audio recording $y(n)$. Note that the genuine speech signal $s(n)$ usually does not have spectrum overlaps with the ENF signal $f(n)$. Therefore, the impact of $s(n)$ can be eliminated after band-pass filtering and the ENF signal can be recovered.

However, in real applications, the audio recording is not guaranteed to be noise free due to the imperfections of recording devices and environments. For this scenario, the speech signal should be formulated by

$$y(n) = s(n) + f(n) + v(n), \tag{2}$$

where $v(n)$ denotes the noise, which is often a broad-band signal. Unfortunately, in most situations, no a priori knowledge about $v(n)$ is available. Therefore, $v(n)$ cannot be completely removed even if a filter with sufficiently narrow pass band is employed. Though there exist some works concerning how to more accurately recover the ENF signal [30–32], yet no further progress has been reported on the issue of ENF extraction under unfavorable noise conditions. In addition, a major problem with these ENF tracking methods is that they require audio recordings of sufficiently long duration. As shown in Fig. 1(a), an obvious ENF component can be clearly observed around 50 Hz in clean speech after band-pass filtering, while this is not the case for noisy speech signals as demonstrated in Fig. 1(b), where the audio recording is subject to unfavorable background noise with a signal to noise ratio (*SNR*) of 15 dB. More severe background noise will further hamper the ENF tracking as shown in Fig. 1(c), where the *SNR* is decreased to 5 dB. By implication, the ENF variations can be concealed by a high level of noise. Hence, detection methods proposed in [20,21,24,25] failed or degraded. In this paper, we move beyond the conventional ENF-based methods to seek for a novel approach to expose tampered noisy speech via spectral phase analysis. The challenges brought by external noise can be overcome through phase reconstruction.

3. STFT analysis and spectral phase recovery

In this work, we consider only the waveform speech with the assumption that the recording keeps its original sampling rate after manipulation. Otherwise, difference in sampling rates between the genuine part and the tampered part can be easily detected by the distinct falloffs since different anti-alias filters are adopted for different sampling rates [21]. We conduct phase recovery on the STFT domain due to the fact that neighboring sub-band phases are highly correlated if the speech signal is transformed to the STFT domain. This correlation between spectral phases can be further used for forgery detection.

3.1. Short-time Fourier transform for speech signals

Let us first recall how a signal can be represented in the STFT domain [33]. For a unified notation, we again use the same symbols as in Section 2 to denote the noisy speech. A noise-free speech can also be formulated by (2), but with $v(n)$ equal to zero. The STFT for noisy speech is represented as

$$Y(k,l) = \sum_{n=0}^{N-1} y(n + l \cdot L) \cdot w(n) \cdot e^{-j\omega_k n} \tag{3}$$

where the noisy speech is processed on a length N basis, in accordance with the fact that audio signals can be viewed as stationary