Digital Signal Processing 45 (2015) 84–95

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Verification of hidden speaker behind transformation disguised voices

Yong Wang [*], Haojun Wu, Jiwu Huang

School of Information Engineering, Wuyi University, No. 22, Dongcheng cun, Jiangmen, Guangdong 529020, China

ARTICLE INFO

Article history:
Available online 26 June 2015

Keywords:
Voice transformation
Disguise
Automatic speaker verification
Security

ABSTRACT

Voice transformation, which has been integrated in many audio (speech) processing tools, is a common operation to change a person's voice and to conceal his or her identity. It can deceive human beings and automatic speaker verification (ASV) systems easily, and thus it presents threats to security. Until now, few efforts have been reported on the recognition of hidden speakers from such disguised voices. In this paper, we propose concrete countermeasures to erase the disguise effects and verify the speaker's identity from voice transformation disguised voices. The proposed system is tested by commonly used audio editors and voice transformation algorithms. The experimental results show that the performances of baseline ASV system without our proposed countermeasures are entirely destroyed by voice transformation disguise with equal error rates (EERs) higher than 40%; while with our proposed countermeasures, the verification performances are improved significantly with EERs lowered to 3%–4%.

1. Introduction

Voice disguise can be classified into two categories: voice conversion (VC) and voice transformation (VT) [1,2]. VC intends to transform one's voice to imitate a target person provided with the target's acoustic information, while VT intends to change the sound without any target. It is apparent that VC is to change one's voice in order "to be recognized as another person" while VT is to change one's voice in order "not to be recognized". Both present threats to human identification. However, since no target information is needed, VT is much easier to implement than VC, leading to the fact that VT has been incorporated in many prevailing audio editors while VC has never been.

Besides electronic disguise by audio editors, VT can be performed by non-electronic means of a mechanic system [3–8] like a mask over the mouth, a pen in the mouth or pinching the nostril. However, by using sophisticated algorithms the output by editors generally sounds much more natural than the mechanic one [1], and thus it presents greater confusion as people tend to be deceived more by voices that sound natural.

With high disguise quality and ease of implementation by abundant tools, digital VT disguise has been used in more and more criminal cases, and has presented threats to security. However, research efforts on this topic is still insufficient. In this paper we will examine automatic speaker verification (ASV) of digital VT disguised voices.

Though few researches on ASV of VT disguise were reported, there have been some related efforts in the past 15 years.

Tan [5], Zhang et al. [6], Perrot et al. [7] and Künzel et al. [8] investigated the effect of several mechanic VT disguises including hand over mouth, pinched nostril, high pitch and low pitch, and conducted experiments using Gaussian Mixture Model (GMM) [9], Vector Quantization (VQ) and Support Vector Machine (SVM) classifiers for ASV. The results indicated that the transformations make the recognition system fail totally. However, no detailed solution has been presented. Jin et al. [10,11], Bonastre et al. [12–14], Wu et al. [15] and Kinnunen et al. [16] studied the effect of VC on ASV systems to find that recognition performance was damaged. Unfortunately, no solution was proposed either. Alegre et al. [17] presented an algorithm based on the reduction in pair-wise distances between consecutive feature vectors, and integrated it into ASV systems. The experimental results showed that with this countermeasure, false acceptance rates (FARs) and equal error rates (EERs) are both significantly lowered compared to without it. However, the authors admitted that this is completely unrealistic of a practical spoofing scenario because the proposed countermeasure exploits prior knowledge of only one limited single spoofing attack. Kons et al. [21] reported an analysis of the vulnerability of text dependent ASV systems to simple VC spoofing attacks. They revealed that training using genuine voices as well as converted voices can improve the verification performance. Like the one in [17], this method also exploits prior full knowledge of only one spoofing attack which is unrealistic in practice; and the method

[*] Corresponding author. Fax: +86 20 39332715.
E-mail addresses: isswy@mail.sysu.edu.cn (Y. Wang), 297049733@qq.com (H. Wu), jwhuang@szu.edu.cn (J. Huang).

http://dx.doi.org/10.1016/j.dsp.2015.06.010
1051-2004/© 2015 Elsevier Inc. All rights reserved.

aims at text dependent ASV, which is far more limited than text independent ASV.

From the above reports, it can be concluded that most of the efforts are focused on VC and mechanic VT disguise, while digital VT disguise has received less attention. More importantly, the major problem of the current work is that most of the researchers only investigate the effects of disguise on ASV systems. They do not present concrete robust and universal solutions to erase the negative effects for revealing genuine speaker's identity, which is a far more challenging issue.

Recently, Wang et al. [18] and Wu et al. [19,20] proposed algorithms to detect VT disguised voices from genuine voices. Cross-disguise-method and cross-corpus were included in the experiments. The results showed that detection rate can reach over 90% in all test situations. These works are significant contributions to VT disguise forensics. But again, a further research is needed on verification of hidden speaker VT disguised voices.

Aiming at the above problems, we will investigate the effects of electronic VT disguise on ASV performances. More importantly, we will propose robust countermeasures for ASV to recognize speaker's identity from disguised voices. We propose a new algorithm for estimating VT parameter estimation by using fundamental frequencies [22,23], and a modified MFCC (Mel Frequency Cepstral Coefficient) extraction algorithm [9] which is effective to recover the original MFCCs from VT disguised voices. The proposed countermeasures are integrated into the GMM–UBM ASV system to test its efforts on the most prevailing and dominant audio editors and algorithms. The proposed system is demonstrated to outperform the baseline system by achieving a low error rate.

The remainder of this paper is organized as follows. In Section 2, we introduce the principles of electronic voice transformations. In Section 3, we propose countermeasures to compensate the deformation of acoustic feature and to erase the disguise effects. Experimental results are given in Section 4. Finally, we summarize conclusions and future works in Section 5.

2. Models of VT and analysis of disguised voices

VT can be divided into two categories: frequency-domain based and time-domain based techniques. In this section we discuss the principle of VT techniques and their deformation effects on acoustic features.

2.1. Model of frequency-domain based VT

The principle of frequency domain based VT is to raise or lower voice pitch by stretching or compressing the frequency spectrum [24]. Being related by Fourier transform, time and frequency characteristics of a signal are not independent but are of a duality relationship. It is essential to break this traditional tie between them to keep the tempo unchanged.

The most frequently used tool for analysis in frequency-domain methods is based on the quasi-stationary sinusoidal model [25], in which speech signal $x(t)$ is represented as a sum of sinusoids whose instantaneous frequency and amplitude vary slowly with time. In most applications, this model is represented by short-time Fourier transform (STFT), which starts with dividing a signal into short segments. Fast Fourier transform (FFT) is then applied to each segment and the resulting spectral components can be manipulated in a variety of ways. However, due to the resolution limitation, FFT bin frequencies generally do not represent true or instantaneous frequencies. For example using a window of size 2048 and a sampling rate of 44.1 kHz, the resolution in frequency domain is only 21.5 Hz, which is far too coarse in the lower frequency band.

In order to solve this problem, a phase-vocoder is introduced which, by insight of relationship between phase and frequency, employs phase information that STFT ignores to improve frequency estimation. The kernel of the phase vocoder is to compute deviation from FFT bin frequencies to instantaneous frequencies by using phase information. Instantaneous frequency can then be computed by adding the deviation and the FFT bin frequency. Finally, three numbers obtained from the FFT analysis for each sinusoid, namely bin magnitude, bin frequency and bin phase, are reduced to only magnitude and transient frequency. We now present it in a simple form in Eqs. (1)–(3).

Suppose $x_t(n)$ is a frame of length N from the input speech signal at time t. Firstly, it is windowed by $w(n)$, and then an FFT is performed on the windowed signal, using Eq. (1), where $w(n)$ is a Hamming or Hanning window and k is the bin frequency index.

$$F(k) = \sum_{n=0}^{N-1} x_t(n) \cdot w(n) e^{-i\frac{2\pi kn}{N}} \quad 0 \leqslant k < N \tag{1}$$

Then, instantaneous magnitude $|F(k)|$ and instantaneous frequency $\omega(k)$ are calculated by Eq. (2) and Eq. (3) respectively,

$$|F(k)| = |\sum_{n=0}^{N-1} x_t(n) \cdot w(n) e^{-i\frac{2\pi kn}{N}}| \quad 0 \leqslant k < N \tag{2}$$

$$\omega(k) = (k + \Delta) \cdot F_s / N \quad 0 \leqslant k < N \tag{3}$$

where F_s is the sampling frequency and Δ is the deviation from the kth bin frequency.

For voice transformation, transient frequency $\omega(k)$ is modified by Eq. (4), where α is the scale factor.

$$\omega'(\lfloor k \cdot \alpha \rfloor) = \omega(k) \cdot \alpha \quad 0 \leqslant k, k \cdot \alpha < N/2 \tag{4}$$

There are several ways to modify the instantaneous magnitude. The commonest method is linear interpolation, as seen in Eq. (5) [24], where $0 \leqslant k, k' < N/2$, $k = \lceil k'/\alpha \rceil$, and $\mu = k'/\alpha - k$.

$$|F'(\lfloor k \cdot \alpha \rfloor)| = \mu |F(k)| + (1 - \mu)|F(k+1)| \tag{5}$$

Another commonly used method is energy-preserving modification by Eq. (6).

$$|F'(\lfloor k \cdot \alpha \rfloor)| = \sum_{\lfloor k \cdot \alpha \rfloor \leqslant k \cdot \alpha < \lfloor k \cdot \alpha \rfloor + 1} |F(k)| \tag{6}$$

For simplicity, we still use k as the index of the modified instantaneous frequency ω' and the instantaneous magnitude F'.

The instantaneous phase $\phi'(k)$ is then calculated via the instantaneous frequency $\omega'(k)$ and the transformed FFT coefficients is obtained by Eq. (7).

$$F'(k) = |F'(k)| e^{i\phi'(k)} \tag{7}$$

An inverse FFT (IFFT) is performed on $F'(k)$ and the transformed signal can thus be obtained.

By phase-vocoder based transformation, all frequency components are adjusted by the scaling factor that includes fundamental frequencies and formants [24]. Considering that fundamental frequencies are more stable and easier to extract than formants, we will use fundamental frequencies to estimate scaling factor.

A phase-vocoder is not universal, however. In some applications, STFT phases are either lost or not applicable with STFT magnitude (STFTM) as the only available information. Hence, algorithms [26–30] have been explored to reconstruct time-domain signal from STFTM without phase information. Among them the latest and most effective algorithms are real-time iterative spectrogram inversion (RTISI) [26] and RTISI with look-ahead (RTISI-LA)