BLIND DETECTION OF ELECTRONIC DISGUISED VOICE

Haojun Wu, Yong Wang, Jiwu Huang

School of Information Science and Technology, Sun Yat-Sen University
Guangzhou, China, 510006

ABSTRACT

Since voice disguise has great negative impact on establishing authenticity of audio evidence in forensics, and has shown an increasing tendency in illegal applications, it is important to identify whether a suspected voice has been disguised or not. However, research on such detection has not been reported. In this paper, we focus on blind detection of electronic disguised voice. Statistical moments of Mel-frequency cepstrum coefficients (MFCC) are extracted as acoustic features of speech signals. Then an approach for detection of disguised voice based on the extracted features and Support Vector Machine (SVM) classifiers is proposed. The extensive experiments demonstrate that detection rates higher than 95% can be achieved, indicating that detection performance of the proposed approach is good.

Index Terms— electronic voice disguise, blind detection, MFCC statistical moments, SVM

1. INTRODUCTION

Voice disguise is an intentional operation to conceal or forge speaker's identity by changing his or her voice tone. Voice disguise methods can be divided into two types [1]: non-electronic and electronic disguise. Non-electronic methods include using falsetto, pinching nostrils, speaking with object in mouth, etc. Electronic disguise is the use of electronic devices to alter voice. Generally by sophisticated algorithms, electronic methods can achieve much more natural disguise performance and present greater confusion on both automatic speaker recognition (ASR) systems and human beings than non-electronic ones. As a result, criminal cases using electronic disguise have been increasing in phone communications, online chatting, and other speech applications in recent years. Hence, detection of electronic disguised voice has become an important and emergent issue. In this paper we research on this topic.

Up to now, according to our best knowledge, research on such detection has not been reported. Several related studies focus on effect of voice disguise on speaker recognition. It is stated in [2], [3] and [4] that disguised voice can degrade speaker recognition performance. Jin et al. [5] investigated voice disguise for speaker de-identification. However, concrete solutions for speaker recognition against voice disguise or detection of disguised voice have not been reported. Hence in our work, we propose an approach for blind detection of electronic disguised voice.

In our proposed approach, statistical moments of Mel-frequency cepstrum coefficients (MFCC) [6] are computed as the acoustic features, and Support Vector Machine (SVM) method is used for classification. The proposed approach is tested by four leading disguise softwares or algorithm, including Cool Edit [7], Audacity [8], PRAAT [9] and real-time iterative spectrogram inversion (RTISI) algorithm [10] [11] in MATLAB. In the experiments, the resulting detection rates achieve good performance of higher than 95% in variable situations.

The rest of this paper is organized as follows. In Section 2, electronic voice disguise is introduced. In Section 3, we propose an approach for detection of disguised voice. In Section 4, experimental results are presented. In Section 5, conclusion and future work are given. Finally in Section 6, we discuss the relation to prior works.

2. ELECTRONIC VOICE DISGUISE

The principle of voice disguise is to raise or to lower voice pitch by stretching or compressing frequency spectrum [12]. In phonetics, pitch is always measured by 12-semitones-division, indicating that pitch can be raised or lowered by 12 semitones at most [13]. A scaling factor of pitch semitones is therefore the disguising factor. Suppose the pitch of a speech frame to be p_0, the disguising factor to be α semitones and the pitch of the modified speech frame to be p, we have [13]:

$$p = 2^{\alpha/12} \cdot p_0 \tag{1}$$

If α is positive, spectrum is stretched and pitch is raised. Otherwise, spectrum is compressed and pitch is lowered. In this paper, a disguising factor $+k$ is used to denote a pitch-raise modification with k semitones, while $-k$ is used to denote a pitch-lower modification with k semitones.

Thanks to 973 Program (2011CB302204) and NSFC (U1135001, 61100168) for funding.

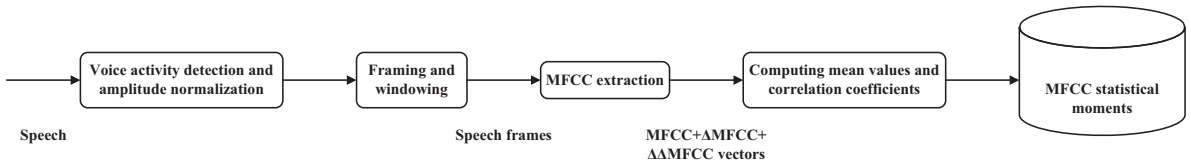

Fig. 1. Extraction procedure of MFCC statistical moments

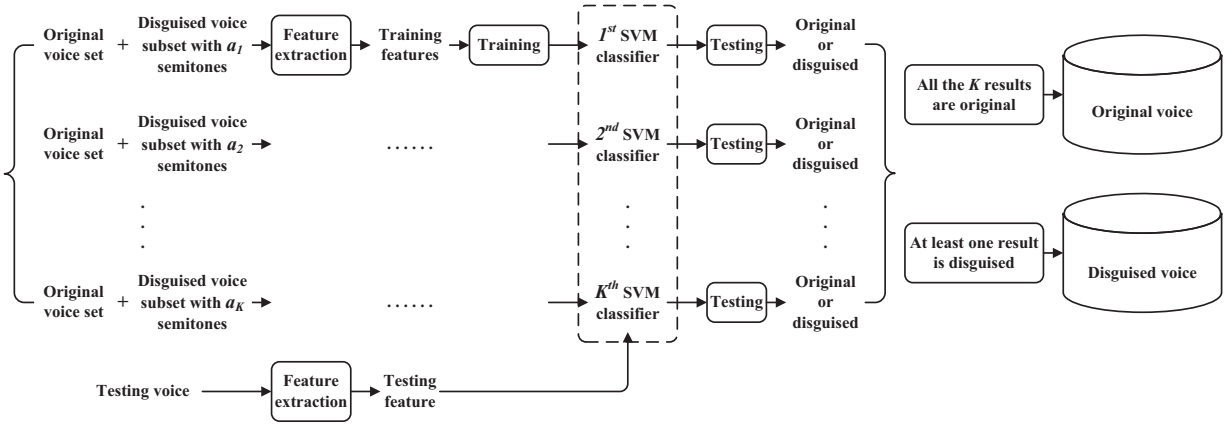

Fig. 2. Generic detection system of disguised voice

3. DETECTION OF DISGUISED VOICE

Since pitch modification on a speech signal changes all the MFCC [14], leaving telltale footprints in statistical moments of MFCC, we use MFCC statistical moments as the acoustic features for modeling in our proposed approach. The detection system is based on SVM, the leading method for pattern classification.

3.1. Feature extraction

Pre-processing including voice activity detection (VAD) and amplitude normalization is performed on a speech signal $s(n)$, followed by framing and hamming windowing to obtain $s'(n)$ with N frames. A d-order MFCC vector is extracted from each frame of $s'(n)$. Then a d-order ΔMFCC vector and a d-order $\Delta\Delta$MFCC vector, which reflect dynamic spectral features, are computed from the MFCC vector. Please refer to [14] for details of MFCC extraction and other related computations.

The above three vectors are concatenated to form a $3d$-order MFCC+ΔMFCC+$\Delta\Delta$MFCC vector of each frame. Suppose v_{ij} to be the jth component of the MFCC+ΔMFCC $+\Delta\Delta$MFCC vector of the ith frame, and V_j to be the set of all the jth components. Then V_j can be expressed as

$$V_j = \{v_{1j}, v_{2j}, \ldots, v_{Nj}\}, j = 1, 2, \ldots, D \tag{2}$$

where $D = 3d$.

In our work, two kinds of vector statistical moments, including mean values of each vector component V_j and correlation coefficients between vector components V_j and $V_{j'}$, are taken into consideration. They are computed by Equ. 3 and Equ. 4, respectively:

$$m_j = E(V_j), j = 1, 2, \ldots, D \tag{3}$$

$$c_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{VAR(V_j)}\sqrt{VAR(V_{j'})}}, j, j' = 1, 2, \ldots, D, j \neq j' \tag{4}$$

Finally, the resulting m_j and $c_{jj'}$ are combined to form MFCC statistical moments of $s(n)$:

$$F = [m_1, m_2, \ldots, m_D, c_{12}, c_{13}, \ldots, c_{D-1D}] \tag{5}$$

F will be used as the final acoustic feature and as the input into SVM classifiers. Extraction procedure of MFCC statistical moments is presented in Fig. 1.

3.2. Detection system

The proposed detection approach is based on MFCC statistical moments and SVM classifiers. The generic detection system of disguised voice is given in Fig. 2.

In training stage, a training database is composed of an original voice set and a disguised voice set. According to K different disguising factors, the disguised voice set is divided into K subsets. MFCC statistical moments are extracted as the acoustic features. Features from the original voice set, with features from each disguised voice subset are used as the training features to train a SVM classifier. Therefore, K SVM classifiers are obtained, and each of them is used to identify whether a testing voice is disguised or not.

In testing stage, the acoustic feature from a testing voice is extracted as the input feature into each resulting SVM classifier. Results of the K SVM classifiers are combined to obtain a final detection result according to the following rule: if all the K results are original, the testing voice is identified as original voice; if at least one result is disguised, the testing voice is identified as disguised voice.

4. EXPERIMENTAL RESULTS

4.1. Experiment setup

TIMIT [15] is used as the corpus in our experiments. It is composed of 6300 speech segments from 630 speakers. The file format is of WAV, 8kHz sampling rate, 16-bit quantization and mono. In our experiments, this corpus is divided into two disjoint parts: TIMIT_1 with 3000 speech segments from 300 speakers for training, and TIMIT_2 with other 3300 speech segments from 330 speakers for testing.

In order to assess the proposed approach comprehensively, four different disguise methods are considered, including Cool Edit, Audacity, PRAAT and RTISI. When disguising factors are too small or too large, disguise performance is not obvious or natural and presents little threaten to ASR systems or human beings. Therefore, we consider 12 kinds of disguised voice with disguising factors from $+4$ to $+9$ semitones and from -4 to -9 semitones.

For a speech signal, 8-order MFCC vectors are extracted.8-order ΔMFCC vectors and 8-order $\Delta\Delta$MFCC vectors are computed from the MFCC vectors. 24-order MFCC+ΔMFCC $+\Delta\Delta$MFCC vectors are then obtained. MFCC statistical moments are then computed by Equ. 3 and Equ. 4 as the acoustic feature of the speech.

Gaussian Radial Basis Function (RBF) kernel and Sequential Minimal Optimization (SMO) method are used in training a SVM classifier [16].

4.2. Detection performance

Two cases are considered in our experiments.

Case 1: Since there are quite many disguise methods, it is very possible that the method used in a criminal case is different from the one for model training. Hence, in this case, we use one disguise method in training stage, and use the four ones in testing stage respectively, which simulates the real forensic scenarios and reveals effect of different disguise methods on the proposed system.

Table 1. Detection performance of disguised voice by different disguise methods

Disguise method for training	Disguise method for testing				FAR
	Cool Edit	Audacity	PRAAT	RTISI	
Cool Edit	99.69%	98.88%	99.48%	99.05%	4.21%
Audacity	96.29%	99.85%	97.90%	96.67%	2.58%
PRAAT	97.55%	97.21%	99.86%	96.58%	2.67%
RTISI	98.65%	97.86%	98.23%	99.69%	4.64%

Detection performance of case 1 is presented in Table 1. It can be seen that when training and testing databases are using the same disguise method, detection rates are higher than 99%. When disguise methods for training and testing are different, detection rates of disguised voice are also steady and higher than 95%. Besides, false alarm rates (FAR) by original voice are lower than 5%. Hence, the proposed approach achieves good detection performance and has strong robustness to different disguise methods.

Case 2: The proposed approach is tested by 12 disguising factors, which reveals effect of disguising factors on detection performance.

In this case, for each disguise method, the disguised voice set is divided into 12 subsets according to the 12 disguising factors. A detection rate is obtained from each subset to represent the degradation degree brought by different disguise degree. Again, one disguise method is used for training, and all the four methods with 12 disguising factors are used for testing.

The result of case 2 in which Cool Edit is used for training is presented in Table 2. It can be seen that detection rates are higher than 95%. With increasing factors, i.e., from $+4$ to $+9$ semitones or from -4 to -9 semitones, the detection rates show an increasing trend, indicating that for a larger factor detection is easier.

The results when Audacity, PRAAT and RTISI are used for training are presented in Table 3, 4 and 5, respectively. Similarly to Table 2, detection rates are higher than 95% for most of the disguising factors. However, when the factor is $+4$ semitones, since disguise performance of these three methods are not obvious enough, several detection rates drop to less then 90%.

It is reasonable to have a performance deterioration with smaller factors. But this is not a simple issue. Instead, serious discussion is needed.

Firstly, as a matter of fact, the details of algorithms adopted by these leading audio processing softwares are not open, i.e., the specific of disguise methods as well as the postprocessing are not known to public. Therefore, it is quite difficult to design a specific detection approach for a specific voice disguise tool without the knowledge of its inner implementation; and this is why we attempt to generate a universal model for all the disguise methods. But again without knowledge of the

Table 2. Detection performance of disguised voice with variable disguising factors. Training set: Cool Edit

Disguise method for testing	Disguising factor (semitones)											
	$+4$	$+5$	$+6$	$+7$	$+8$	$+9$	-4	-5	-6	-7	-8	-9
Cool Edit	97.18%	99.58%	99.76%	99.94%	99.94%	99.97%	99.97%	100.00%	100.00%	100.00%	100.00%	100.00%
Audacity	95.48%	99.15%	99.85%	100.00%	99.97%	100.00%	95.52%	97.55%	99.09%	99.97%	100.00%	100.00%
PRAAT	95.85%	99.48%	99.88%	99.97%	99.97%	99.97%	98.61%	100.00%	100.00%	100.00%	100.00%	100.00%
RTISI	91.52%	98.21%	99.55%	99.88%	100.00%	100.00%	99.42%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 3. Detection performance of disguised voice with variable disguising factors. Training set: Audacity

Disguise method for testing	Disguising factor (semitones)											
	$+4$	$+5$	$+6$	$+7$	$+8$	$+9$	-4	-5	-6	-7	-8	-9
Cool Edit	75.39%	90.00%	95.18%	97.30%	98.21%	99.39%	99.94%	100.00%	100.00%	100.00%	100.00%	100.00%
Audacity	98.73%	99.52%	99.94%	100.00%	100.00%	100.00%	99.97%	100.00%	100.00%	100.00%	100.00%	100.00%
PRAAT	83.15%	95.67%	98.70%	99.82%	99.79%	99.82%	98.00%	100.00%	100.00%	99.97%	99.97%	99.97%
RTISI	74.30%	90.73%	97.24%	99.30%	99.70%	99.85%	98.97%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 4. Detection performance of disguised voice with variable disguising factors. Training set: PRAAT

Disguise method for testing	Disguising factor (semitones)											
	$+4$	$+5$	$+6$	$+7$	$+8$	$+9$	-4	-5	-6	-7	-8	-9
Cool Edit	84.42%	93.85%	96.55%	98.03%	98.30%	99.45%	100.00%	100.00%	100.00%	100.00%	100.00%	99.97%
Audacity	90.00%	96.70%	99.00%	99.52%	99.67%	99.85%	91.91%	93.88%	97.39%	99.18%	99.67%	99.73%
PRAAT	98.58%	99.82%	99.91%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
RTISI	74.48%	90.64%	96.42%	98.58%	99.21%	99.64%	100.00%	100.00%	100.00%	100.00%	99.97%	99.97%

Table 5. Detection performance of disguised voice with variable disguising factors. Training set: RTISI

Disguise method for testing	Disguising factor (semitones)											
	$+4$	$+5$	$+6$	$+7$	$+8$	$+9$	-4	-5	-6	-7	-8	-9
Cool Edit	89.64%	97.24%	98.52%	99.33%	99.33%	99.79%	99.97%	100.00%	100.00%	100.00%	100.00%	100.00%
Audacity	94.42%	98.52%	99.64%	99.97%	99.94%	99.97%	90.76%	93.97%	97.91%	99.52%	99.82%	99.88%
PRAAT	84.58%	96.03%	99.03%	99.61%	99.85%	99.91%	99.91%	100.00%	100.00%	99.97%	99.97%	99.97%
RTISI	97.21%	99.45%	99.73%	99.97%	99.97%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

specific algorithms, it becomes difficult to conduct a theoretical analysis for modeling defects. To know about the inner details and to research on a better modeling and classification method are our future works if possible.

Secondly, it is commonly known that some biometrics like fingerprints and iris have been widely used and accepted by detectives, lawyers, judges and law enforcement agencies for forensic purposes. However, due to unstable characteristics of speech signals and the ease of modification of digital audio without footprints, digital audio forensics has not been considered as solid as others and has not been accepted universally. From a practical point of view, we think that although the present research on audio forensics is not thought to yield solid evidence, it can help in forensics and security to some extend like narrowing the scope of suspects or early warning of suspected voice. From this point of view, the overall performance of our proposed approach is quite good considering that most of the voice segments can be identified correctly even when they are disguised naturally. We can also observe that when Cool Edit is used for training the overall performance is the best and excellent. This is a quite satisfactory observation because Cool Edit is the most prevailing audio processing software. It is quite easy to obtain abundant speech signals processed by Cool Edit, indicating that we can have abundant materials for training models.

5. CONCLUSION

In this paper, an approach for blind detection of electronic disguised voice is proposed. MFCC statistical moments are extracted as acoustic features of speech signals and SVM is used as classification method. In the experiments, four kinds of commonly used disguise methods are used for training and testing. Detection performance of the proposed approach is demonstrated to be excellent, even when disguise methods used in training stage are different from the ones used in testing stage. However, when the disguising factor is $+4$ semitones, the detection performance is not good enough. It is our future work to analyse the reason of such situation and to improve the detection performance.

6. RELATION TO PRIOR WORK

The work presented in this paper focuses on blind detection of electronic disguised voice. Research on such detection has not been reported. The related studies include the work by Tan [2] discussing effect of disguised voice on speaker recognition, and the work by Jin et al. [5] studying using voice disguise for speaker de-identification. Detection of disguised voice is not considered in these earlier studies, while it is studied in our present work.

7. REFERENCES

[1] R.D. Rodman, "Speaker recognition of disguised voices: a program for research", Department of Computer Science, North Carolina State University, Raleigh, North Carolina, 2003.

[2] Tiejun Tan, "The effect of voice disguise on automatic speaker recognition", in *Congress on Image and Signal Processing, CISP*, 2010.

[3] Cuiling Zhang and Tiejun Tan, "Voice disguise and automatic speaker recognition", in *Forensic Science International, FORENSIC SCI INT*, 2008, vol. 175, no. 2, pp. 118–122.

[4] H.J. Kunzel, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Effect of voice disguise on the performance of a forensic automatic speaker recognition system", in *ODYS*, 2004, pp. 153–156.

[5] Qin Jin, A.R. Toth, T. Schultz, and A.W. Black, "Voice convergin: speaker de-identification by voice transformation", in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2009, pp. 3909–3912.

[6] D.A. Reynolds and R.C. Rose, "Robust text independent speaker identification using gaussian mixture speaker models", in *IEEE Transactions on Speech and Audio Processing, IEEE SAP*, 1995, vol. 3, no. 1, pp. 72–83.

[7] "Cool edit pro is now adobe audition [online]", in *http://www.adobe.com/products/audition.html*.

[8] "Audacity: free audio editor and recorder [online]", in *http://audacity.sourceforge.net*.

[9] "Praat: doing phonetics by computer [online]", in *http://www.fon.hum.uva.nl/praat*.

[10] "Time-scale/pitch modification tools for matlab [online]", in *http://www.mathworks.com/matlabcentral/fileexchange/25880-time-scalepitch-modification*.

[11] Xinglei Zhu, G.T. Beauregard, and L.L. Wyse, "Real time signal estimation from modified short time fourier transform magnitude spectra", in *IEEE Transactions on Audio, Speech and Language Processing, TASLP*, 2007, vol. 15, no. 5, pp. 1645–1653.

[12] J. Laroche, "Time and pitch scale modification of audio signals", Joint E-mu/Creative Technology Center.

[13] S.E. Trehub, A.J. Cohen, L.A. Thorpe, and B.A. Morrongiello, "Development of the perception of musical relations: semitone and diatonic structure", in *Journal of Experimental Psychology-human Perception and Performance, J EXP PSYCHOL-HUM PERCEP PERF*, 1986, vol. 12, no. 3, pp. 295–301.

[14] Wei Han, Cheong-Fat Chan, Oliver Chiu-Sing Choy, and Kong-Pang Pun, "An efficient mfcc extraction method in speech recognition", in *IEEE International Symposium on Circuits and Systems, ISCAS*, 2006.

[15] "Timit acoustic-phonetic continuous speech corpus [online]", in *http://www.ldc.upenn.edu/Catalog/ CatalogEntry.jsp?catalogId=LDC93S1*.

[16] "Libsvm tool [online]", in *http://www.csie.ntu.edu.tw/cjlin/libsvm*.