# Blind Detection of Electronic Voice Transformation with Natural Disguise

Yong Wang, Yanhong Deng, Haojun Wu, and Jiwu Huang

School of Information Science and Technology, Sun Yat-sen University, Higher Education Mega Center, PanYu, Guangzhou 510006, China
isswy@mail.sysu.edu.cn

**Abstract.** Electronic voice transformation with natural disguise ability is a common operation to change a person's voice and conceal his or her identity, which can easily cheat human ears and automatic speaker recognition(ASR) systems and thus presents threaten to security. Till now, few efforts have been reported on detection of electronic transformation, which aims to distinguish disguised voices from original voices. Therefore in this paper we investigate the principle of electronic voice transformation, and propose a blind detection approach using MFCC(Mel Frequency Cepstrum Coefficients) as the acoustic features and VQ-SVM (Vector Quantization-Support Vector Machine) as the classification method. By extensive experiments, it is demonstrated to have classification accuracy higher than 98% in most cases, indicating that the proposed approach has good performance and can be used in forensic applications.

**Keywords:** Electronic voice transformation, disguise, detection, forensic.

## 1 Introduction

Voice disguise can be classified into two categories: voice conversion and voice transformation[1][2]. Voice conversion is to transform one's voice to imitate a target person provided with the target's acoustic information, while voice transformation is to change the sound without any target. Both conceal speaker's identity and present threaten to security. However, since no target information is needed, voice transformation is much easier to implement and be adopted in criminal cases than voice conversion. Therefore, in this paper we will focus on detection of voice transformation disguise.

Voice transformation can be implemented by non-electronic and electronic means. Non-electronic means includes the alteration of voice by using a mechanic system like a mask over the mouth, a pen in the mouth or pinching the nostril. Electronic means is achieved by softwares or electronic devices. Generally, by sophisticated algorithm the electronic output sounds more natural than the non-electronic one and presents greater confusion, since people may be easier to be cheated by transformed voices that sound natural. Moreover, electronic

Y.Q. Shi, H.J. Kim, and F. Pérez-González (Eds.): IWDW 2012, LNCS 7809, pp. 336–343, 2013.

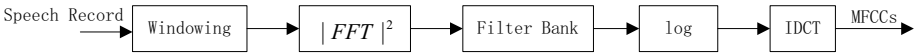


**Fig. 1.** MFCC extraction process

voice transformation has been incorporated into many softwares and electronic devices in recent years, and has been adopted in more criminal cases than before. However, research efforts on forensic of electronic voice transformation is still very insufficient. Hence in this paper we will examine detection of electronic voice transformation, which can be used in many forensic applications, especially in those related to speaker recognition. For example in forensic speaker recognition(FSR), the detection results can be used a a reference to aid the recognition decision making.

There are several voice transformation methods. Since our aim is to investigate the one with natural disguise ability that presents threaten both to human ears and to automatic speaker recognition(ASR) systems, we will focus on phase vocoder based transformation [3] because it presents the most natural disguise ability among the existing voice transformation methods, and it is prevailing.

Time and frequency characteristics of a signal, being related by Fourier transform, are not independent but of a duality relationship. Since 1950s, large collection of voice transformation solutions have been proposed which break this tradition tie between pitch and rate of playback. Among them, the most frequently used techniques are based on 'signal models' [4] and implemented by short-time Fourier transform(STFT), also referred to as phase-vocoder, that change the prosody by pitch modification[3]. The phase-vocoder based method has been incorporated in many professional and popular softwares for speech and music processing. One of such leading benchmarks is Adobe Audition[5]. In this paper we will examine the blind detection of phase vocoder prototype transformation[6] and further verify our proposed approach by its implementation in Adobe Audition.

Most of the adjacent existing researches focus on investing effects of transformation on ASR systems [1][7][8][9][10][11][12]. However, there have been no reported research efforts on blind detection of voice transformation till now. Therefore in this paper, based on the study of voice transformation principle, we will propose a blind and robust detection approach to distinguish disguised voices from original voices using MFCCs(Mel Frequency Cepstrum Coefficients)[9] as the acoustic features and VQ-SVM(Vector Quantization - Support Vector Machine)[13][14] as the classification method.

The structure of this paper is as follows. In Section 2 we introduce the model of electronic voice transformation and MFCC extraction process. In Section 3 we present our proposed blind detection approach using MFCC and VQ-SVM. Experimental results are given in Section 4. Finally we summarize conclusions and future works in Section 5.

## 2 Model of Electronic Voice Transformation and MFCC Extraction

In applications of audio processing, the most widely used analysis/synthesis method is short-time Fourier transform(STFT) which begins by windowing a signal to 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, the FFT bin frequencies generally do not represent the true frequencies(also called instantaneous frequencies). For example using a window of size 2048 and a sampling rate of 44.1kHz, the resolution in frequency domain is only 21.5Hz, which is far too coarse in lower frequency band.

By insight of the relationship between phase and frequency, phase vocoder employs phase information that the STFT ignores to improve frequency estimation. The core of phase vocoder is to compute the deviation from the FFT bin frequency to the instantaneous frequency 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 just magnitude and transient frequency. The entire procedure can be referred to [3]. We now present it in a simple form in Equ.(1)-(3).

Firstly speech signal $x(n)$ is windowed by hamming or hanning window by Equ.(1).

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

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

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

$$\omega(k) = (k + \Delta) * Fs/N \tag{3}$$

where $Fs$ is the sampling frequency and $\Delta$ is the deviation from the $k^{th}$ bin frequency.

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

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

In order to maintain the signal energy, transient magnitude is modified by Equ.(5)

$$|F'(\lfloor k * \alpha \rfloor)| = \sum_{\lfloor k*\alpha \rfloor \leqslant k*\alpha < \lfloor k*\alpha \rfloor + 1} |F(k)| \tag{5}$$

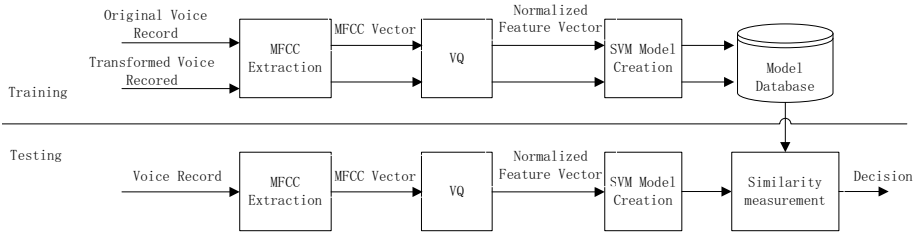


**Fig. 2.** Generic system using MFCCs as the feature and VQ-SVM as the classification method

The instantaneous phase $\phi'(k)$ is then calculated via the instantaneous frequency $\omega'(k)$ and the transformed FFT coefficients is obtained by Equ.(6).

$$F'(k) = |F'(k)|e^{j\phi'(k)} \tag{6}$$

Inverse FFT is performed on $F'(k)$ and the transformed signal can be obtained.

From the above transformation procedure, we know that the distribution of spectra magnitude $|F(k)|$ is changed after the transformation. Since the voice spectrum carries the natural acoustic features of vocal tract, we think that voice transformation would alter such natural features, and thus the magnitude $|F(k)|$ can be used to distinguish transformed voices from original voices. However, if $|F(k)|$ is used, the computation load will be so heavy that the whole classification process will be extremely slow. Hence we turn to MFCCs which converts the spectrum magnitudes into mel-frequency cepstrum with much less feature data, and which maintains the vocal tract acoustic features.

MFCC extraction process is shown in Fig.1. We can learn that MFCC is computed from spectrum magnitudes $|F(k)|$. The filter bank is to combine the spectrum magnitudes in each mel subband to form one contribution that reflects the vocal tract acoustic features. Furthermore the MFCC extraction algorithm compresses the data amount to shorten the classification process. In this paper a 12-order MFCC vector is extracted from each frame and the whole feature is the combination of MFCC vectors from all the frames.

## 3 Classification Algorithm

Durations of different speech signals are usually different. Since MFCCs are extracted along the time axis, the lengths of MFCC vectors, i.e., the numbers of MFCC vectors from different speech signals are usually different too. However, SVM requires the input features to have the same length. Hence we use VQ(vector quantization) to normalize their lengths.

Suppose $X = \{x_1, x_2, \cdots, x_T\}$ is the MFCCs consisting of $T$ number of vectors extracted from $T$ frames, each of which has dimension $N$. VQ aims to convert $X$ into $K$ separate clusters($K \ll T$). Each cluster is represented by a

code vector, $c_i$, which is the centroid of the cluster. The whole set of code vectors $C = \{c_1, c_2, ..., c_K\}$ is used as the normalized feature vectors input into SVM.

SVM has become the state-of-the-art classification method in many pattern recognition applications in recent years. In most applications SVM is a binary classifier which models the decision boundary in hyper-plane with the maximum margin of separation between two classes. Reported efforts have shown that the best results for SVMs in speech applications have been obtained using the generalized linear discriminant sequence (GLDS) kernel[15][16], which creates a single characteristic vector using the sequence of features extracted from a speech signal. Hence in our SVM configuration GLDS is used as the kernel.

Generic classification system is shown in Fig.2. In training phase, MFCCs of original and transformed speech records are extracted and input into VQ box to normalize the lengths, then the normalized feature vectors are input into SVM to create the models of original and transformed speech records. In testing phase, the same process is carried on the test speech record to create its model and to calculate the similarity between the test speech record and the two models obtained in training phase. Finally a decision is made.

## 4 Experiments

### 4.1 Experiment Setup

1. Corpus
   TIMIT[17] is used as the corpus for the experiments. TIMIT is a popular corpus which has been designed for the development of automatic speech recognition systems and can be used in many other speech applications. It contains 630 speakers, 192 females and 438 males, from 8 major dialect regions of America. Each speaker reads 10 sentences to obtain a total number of 6300 records in it. All the sentences are recorded as as wav format of 16kHz sampling rate, 16-bit quantization and mono channel. The duration of each record lasts around 3 seconds.
2. Transformation methods
   Prototype of phase vocoder transformation[6] and its implementation in Adobe Audition are performed on TIMIT records to obtained transformed records. The transformation scaling factors are $[0.55 : 0.05 : 1.45]$ which cover the range that is commonly used in real applications. Scaling factor 1.0 is excluded since it does not modify the records.
3. Training and testing configuration
   In order to simulate the real forensic scenario, we permute the 6300 records, and divide them in half into a training subset and a testing subset each of which contains 3150 records. Each record in the training subset is transformed with scaling factors $[0.55 : 0.05 : 1.45]$ to obtained 18 transformed records. Original and transformed records in the training subset are then input into VQ-SVM to train the two models. The resulting models are used in testing phase where records in testing subset are also transformed, and

**Table 1.** Classification results of scenario 1 in case of phase vocoder prototype

| scaling factor | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|
| TP | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 98.79% | 93.81% |
| TN | 100% | 100% | 100% | 100% | 100% | 100% | 99.94% | 99.37% | 96.03% |
| accuracy | 100% | 100% | 100% | 100% | 100% | 100% | 99.97% | 99.08% | 94.92% |
| scaling factor | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 1.30 | 1.35 | 1.40 | 1.45 |
| TP | 95.62% | 96.63% | 97.94% | 98.48% | 98.79% | 99.27% | 99.49% | 99.49% | 99.65% |
| TN | 97.37% | 98.29% | 98.92% | 99.40% | 99.62% | 99.78% | 99.81% | 99.90% | 99.90% |
| accuracy | 96.50% | 97.46% | 98.43% | 98.99% | 99.21% | 99.53% | 99.65% | 99.70% | 99.78% |

**Table 2.** Classification results of scenario 1 in case of Adobe Audition

| scaling factor | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|
| TP | 99.65% | 99.46% | 99.37% | 98.89% | 98.48% | 97.30% | 66.70% | 88.13% | 72.16% |
| TN | 99.81% | 99.81% | 99.37% | 98.95% | 97.81% | 97.08% | 73.21% | 86.60% | 73.08% |
| accuracy | 99.73% | 99.64% | 99.37% | 98.92% | 98.15% | 97.19% | 69.96% | 87.37% | 72.62% |
| scaling factor | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 1.30 | 1.35 | 1.40 | 1.45 |
| TP | 58.63% | 84.98% | 79.97% | 100% | 100% | 100% | 99.97% | 100% | 100% |
| TN | 65.08% | 89.33% | 99.05% | 99.97% | 100% | 100% | 100% | 100% | 100% |
| accuracy | 61.86% | 87.16% | 89.51% | 99.99% | 100% | 100% | 99.99% | 100% | 100% |

are input into VQ-SVM along with the original records. SVM outputs *accuracy*, the true classification rate, *TP*, the true classification rate for identifying transformed records, and *TN*, the true classification rate for identifying original records.

### 4.2 Classification Performance

For a thorough investigation we take into consideration the following two scenarios.

1) Scenario 1: Classification between the original records and the transformed records with one scaling factor, which reveals the different effects brought by different transformation degrees.

2) Scenario 2: Classification between the original records and all of the transformed records, which yields the overall performance.

Detection results of scenario 1 in case of phase vocoder prototype transformation is given in Table 1, from which we can have the following observations.

1) The farther the scaling factor is from 1.0, the higher accuracy, TP and TN are achieved, which coincides the fact that the more the records are modified the more difference will be introduced between original and transformed records.

2) TP, TN, and accuracy with scaling factors [0.95,1.05] are apparently much lower than those with other scaling factors, which is considered to be acceptable in this paper, since in real applications scaling factors in range of [0.95 1.05] are of small modification, and they generally does not change the speaker's identity.

3) With scaling factors outside [0.95 1.05] most of the accuracy, TP and TN are over 97% to 100%, indicating the proposed detection approach can distinguish the transformed and original records at a high classification rate.

Detection results in scenario 2 in case of phase vocoder prototype transformation are TP = 95.81%, TN = 98.41%, and accuracy = 97.11%. Since there has been no reported research efforts on such detection, comparison and evaluation are quite difficult to make. However, since the detection results are to be used as reference for the FSR decision making and a FSR system generally yields a much lower accuracy[18], we think that the performance of our proposed approach is acceptable.

Detection results of scenario 1 in case of Adobe Audition is given in Table 2, from which we can see that with scaling factors between [0.85 1.15], TP, TN and accuracy are much lower than the counterparts in case of phase vocoder prototype. They further lower the classification rates in scenario 2: TP = 75.37%, TN = 94.83%, and accuracy = 85.10%. Although detailed algorithms used in Adobe Audition are not open sources, by perception of the acoustic quality, we find that with scaling factors between [0.85 1.15], the signal transformed by Adobe Audition sounds smoother than the one transformed by the prototype. By observation of the frequency domain, we also notice that there are fewer abrupt spectral components in the former than in the latter. This smoothing post-processing by Adobe Audition removes those abrupt characteristics that are not produced by human vocal tract, and makes a transformed record sound more closer to the one that are uttered by vocal tract. With a larger transformation degree, such smoothness declines and the detection rates become similar to the counterparts in case of prototype.

## 5 Conclusions

In summary, in this paper the principle of voice electronic transformation is thoroughly investigated, and a blind detection using MFCCs as the acoustic features and VQ-SVM as the classification approach is proposed to distinguish disguised voices form original voices. Extensive experiments are conducted. The results show that in most cases high TP, TN and accuracy are achieved indicating that the detection results can be used as reference for forensic works. There are still many open issues worthy of investigation in future works. We will continue to improve the performance of our proposed approach, and trace other disguise methods with natural ability if they emerge.

**Acknowledgements.** This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 61100168, the Research Fund for the Doctoral Program of Higher Education of China under Grant 20110171120052, the Fundamental Research Funds for the Central Universities under Grants 12lgpy38, and the National Natural Science Foundation of China (NSFC) under Grants F020703.

## References


1. Perrot, P., Aversano, G., Chollet, G.: Voice disguise and automatic detection: Review and perspectives. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) WNSP 2005. LNCS, vol. 4391, pp. 101–117. Springer, Heidelberg (2007)
2. Masthoff, H.: A report on voice disguise experiment. Forensice Linguistics 3(1), 160–167 (1996)
3. Laroche, J.: Time and Pitch Scale Modification of Audio Signals. In: Applications of Digital Signal Processing to Audio and Acoustics in the International Series in Engineering and Computer Science, vol. 437, ch. 7, pp. 279–309. Springer, US (2006)
4. McAulay, R., Quatieri, T.: Speech analysis/Synthesis based on a sinusoidal representation. IEEE Trans. on Acoustics, Speech and Signal Processing 34(4), 744–754 (1986)
5. `http://www.adobe.com/products/audition.html`
6. Phase vocoder based voice transformation source code, `http://www.dspdimension.com/smbPitchShift.cpp`
7. Perrot, P., Chollet, G.: The question of disguised voice. The Journal of the Acoustical Society of America 123(5), 3878 (2008)
8. Künzel, H., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: Effect of voice disguise on the performance of a forensic automatic speaker recoginition system. In: Proc. of Odyssey, pp. 153–156 (2004)
9. Reynolds, D., Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Processing 3(1), 72–83 (1995)
10. Bonastre, J.-F., Matrouf, D., Fredouille, C.: Artificial impostor voice transformation effects on false acceptance rates. In: Proc. of Interspeech, pp. 2053–2056 (2007)
11. Matrouf, D., Bonastre, J.-F., Costa, J.P.: Effect of impostor speech transformation on automatic speaker recognition. In: Proc. of COST275 Workshop "Biometric on the internet", Hatfied, UK (2005)
12. Matrouf, D., Bonastre, J.-F., Fredouille, C.: Effect of voice transformation on impostor acceptance. In: Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, Toulouse, France (2006)
13. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Trans. Commun. 28, 84–94 (1980)
14. Libsvm Tool, `http://www.csie.ntu.edu.tw/cjlin/libsvm`
15. Campbell, W.M., Campbell, J., Reynolds, D.A., Torres-Carrasquillo, P.: Support vector machines for speaker and language recognition. Computer Speech and Language 20(2), 210–229 (2006)
16. Campbell, W.M., Singer, E., Torres-Carrasquillo, P.A., Reynolds, D.A.: Language recognition with support vector machines. In: Proc. Odyssey: The Speaker and Language Recognition Workshop, pp. 41–44 (2004)
17. TIMIT Acoustic-Phonetic Continuous Speech Corpus, `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`
18. Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.-F., Matrouf, D.: Forensic speaker recognition. IEEE Signal Processing Magazine 26(2), 95–103 (2009)