Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition

Zhizheng Wu[1], *Eng Siong Chng*[1], *Haizhou Li*[1,2,3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Human Language Technology Department, Institute for Infocomm Research , Singapore
[3]School of EE & Telecom, University of New South Wales, Australia

wuzz@ntu.edu.sg, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

Abstract

Voice conversion techniques present a threat to speaker verification systems. To enhance the security of speaker verification systems, We study how to automatically distinguish natural speech and synthetic/converted speech. Motivated by the research on phase spectrum in speech perception, in this study, we propose to use features derived from phase spectrum to detect converted speech. The features are tested under three different training situations of the converted speech detector: a) only Gaussian mixture model (GMM) based converted speech data are available; b) only unit-selection based converted speech data are available; c) no converted speech data are available for training converted speech model. Experiments conducted on the National Institute of Standards and Technology (NIST) 2006 speaker recognition evaluation (SRE) corpus show that the performance of the features derived from phase spectrum outperform the mel-frequency cepstral coefficients (MFCCs) tremendously: even without converted speech for training, the equal error rate (EER) is reduced from 20.20% of MFCCs to 2.35%.

Index Terms: Speaker verification, voice conversion, anti-spoofing attack, synthetic speech detection, phase spectrum

1. Introduction

Speaker verification is a process to confirm a claim of identity based on the user's speech samples [1, 2]. There are two possible outcomes: the claim is accepted or rejected. On the other hand, voice conversion is to modify one speaker's voice (source speaker) to sound like that of another speaker (target speaker) [3, 4, 5]. Therefore, voice conversion technique can be used to modify an impostor's voice to sound like that of the claimed speaker to attack the speaker verification system.

In response to such a potential threat, many studies have been conducted on the vulnerability of speaker verification systems against synthetic speech. The security or vulnerability of speaker verification systems have been studied against imposture using synthetic speech from HMM-based speech synthesis system [6, 7, 8], adapted statistical speech synthesis system [9] and voice conversion techniques [10, 11], which are carried out on high quality speech.

Spoofing attack researches on telephone speech are also conducted. In [12], The vulnerability of GMM-based speaker verification system against voice conversion attacks is studied on the National Institute of Standards and Technology (NIST) 2006 speaker recognition evaluation (SRE) corpus. In our previous works [13], the performance of five speaker verification systems including the state-of-the-art joint factor analysis (JFA) system against spoofing attacks using GMM-based voice conversion techniques are evaluated on NIST 2006 SRE corpus. Then the vulnerability of the JFA system against unit-selection based and GMM-based converted speech is also compared on NIST 2006 SRE corpus [14].

All the previous studies on both high quality speech and telephone speech confirmed that synthetic speech from statistical speech synthesis systems or voice conversion techniques present a threat to speaker verification systems.

In [15], a technique was studied that makes use of the differences in the relative phase shift [16] between high quality human and synthetic speech signals to detect synthetic speech from a HMM-based speech synthesis system. In this study, it was assumed that the synthetic speech data are available for training a detector. However, in real applications, we don't have the information as to what kinds of speech synthesis techniques or voice conversion techniques are used in the spoofing attacks. Hence, no synthetic speech data are actually available. In this study, we focus on detecting spoofing attacks using voice conversion techniques on telephone speech, and propose to use features derived from the phase spectrum, which has been shown to be useful in speech recognition [17, 18], to detect converted speech without the need to know the exact voice conversion techniques.

2. Phase information in voice conversion

Although converted speech has been shown to be able to confuse a speaker verification system in many reported work [10, 11, 13, 14], informal listening tests show that human ear can easily distinguish natural speech and converted speech. It has been shown that phase spectrum is useful for speech perception [19, 20] , and the fact that the original/natural phase information is missing in the converted speech, we would like to look into how converted speech can be detected by using phase features.

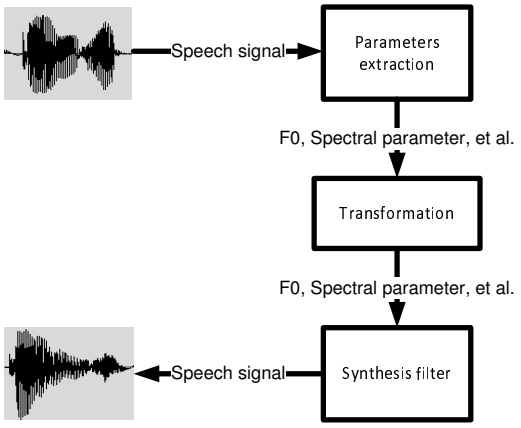

Figure 1: *Typical voice conversion procedure.*

In Figure 1, we show a typical voice transformation procedure. A speech signal is first analyzed to extract parameters, such as fundamental frequency, spectral envelop parameters, which are then converted by an already trained transformation function. Finally the converted parameters are passed to synthesis filter to reconstruct waveform. In the parameter extraction step, we usually extract fundamental frequency and speech envelop parameters, ignoring the phase information. As a result, the original/natural phase information is not kept in the reconstructed waveform.

The best way to build a converted speech detector is to train a converted speech model that characterizes the signal. If converted speech data are available, converted speech models can be trained right away. One notes that the difference between voice conversion techniques lies in the transformation module in Figure 1. In the case when converted speech data are not available, if we can extract the parameters from natural speech and pass directly to the synthesis filter, it is possible that the resulting synthesized speech carry the characteristics of converted speech. We will look into how such synthesized speech can be used to train the converted speech model.

3. Phase information extraction for modeling

Although speech signal is non-stationary, it can be assumed as quasi-stationary. Therefore, speech signal can be processed by short-time Fourier analysis. Given a speech signal $x(n)$, the short-time Fourier transform (STFT) is given as follows:

$$X(\omega) =\mid X(\omega) \mid e^{j\psi(\omega)} \tag{1}$$

where $\mid X(\omega) \mid$ is the short-time magnitude spectrum, and $\psi(\omega)$ is the short-time phase spectrum. The square of magnitude spectrum, $\mid X(\omega) \mid^2$, is usually called the power spectrum, from which the mel-frequency cepstral coefficients (MFCCs) are derived.

As natural phase information is missing in the reconstructed speech/converted speech, phase spectrum can be used to derive features providing the evidence of converted speech. In this study, two different features will be derived from phase spectrum as follows.

3.1. Cosine normalization of phase spectrum

As the original phase spectrum is not continuous in frequency domain, we first unwrap the phase spectrum as a continuous function of frequency. After unwrapping, the range of spectrum could vary which makes it difficult to model phase information. Cosine function is then applied on the unwrapped phase spectrum to normalize the range into $[-1.0, 1.0]$. Then discrete cosine transform (DCT) is applied on the Cosine normalized phase spectrum to reduce dimensionality. Finally we keep 12 cepstral coefficients, excluding the 0th coefficient. This feature is called *cos-phase* in this study.

3.2. Frequency derivative of phase spectrum

The second feature is the frequency derivative of phase spectrum, obtained by the group delay function (GDF) which is a measure of the nonlinearity of the phase spectrum [17] and defined as the negative derivative of the phase spectrum with respect to ω:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{\mid X(\omega) \mid^2} \tag{2}$$

where $Y(\omega)$ is the STFT of $nx(n)$; $X_R(\omega)$, $X_I(\omega)$ and $Y_R(\omega)$, $Y_I(\omega)$ are the real part and imaginary part of $X(\omega)$ and $Y(\omega)$, respectively.

To capture the fine structure of group delay phase spectrum, in practice, a modified group delay function is adopted. Smoothed power spectrum is used instead of original power spectrum and two variables are introduced to emphasis the fine structure of the phase spectrum. The modified group delay function is described as follows.

$$\tau_\gamma(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{\mid S(\omega) \mid^{2\gamma}} \tag{3}$$

$$\tau_{\alpha,\gamma}(\omega) = \frac{\tau_\gamma(\omega)}{\mid \tau_\gamma(\omega) \mid} \mid \tau_\gamma(\omega) \mid^\alpha \quad (4)$$

where $\mid S(\omega) \mid^2$ is the smoothed power spectrum, $\tau_{\alpha,\gamma}(\omega)$ is the modified group delay phase spectrum, α and γ are two variables to make the phase spectrum be presented in favorable form, α and γ are set to 0.4 and 1.2 in this study, respectively.

After modified group delay phase spectrum is obtained, discrete cosine transform (DCT) is applied. We keep 12 cepstral coefficient, excluding the 0th coefficient. The 12 dimension feature are used for model training and detection and this feature is called *MGDF-phase* in this study.

4. Experimental setups and results

In this study, the performance of cos-phase and MGDF-phase features (without delta and delta-delta coefficients) are evaluated for converted speech detection under three different training situations: a) only GMM-based converted speech data are available; b) only unit-selection based converted speech data are available; c) no converted speech data are available. We use 12 dimension mel-frequency cepstral coefficients (MFCCs) and their delta and delta-delta coefficients, which are used in speaker verification [13, 14], as the reference baseline.

The natural/converted speech decision is made using log-scale likelihood ratio as follows:

$$\Lambda(C) = \log p(C|\lambda_{converted}) - \log p(C|\lambda_{natural}) \quad (5)$$

where C is the feature vector sequence of a speech signal, $\lambda_{converted}$ is the GMM model for converted speech, and $\lambda_{natural}$ is the GMM model for natural speech. Under the three different situations, we have the same natural speech model $\lambda_{natural}$, but three different converted speech model $\lambda_{converted}$. The number of Gaussian components of GMM is set to 512. Equal error rate (EER) is reported as the evaluation criterion.

4.1. Corpora

A subset of the NIST 2006 SRE corpus, and a subset of the spoofing attack corpora are adopted. The spoofing attack corpora are created using two different voice conversion methods (GMM-based voice conversion and unit-selection based voice conversion method) in our previous studies [13, 14]. In constructing the converted speech corpora, mel-cepstral analysis [21] and MLSA filter [21] are adopted for feature extraction and waveform reconstruction, respectively. In practice, the Speech Signal Processing Toolkit (SPTK) tool [22] is used.

We use 100 sessions of natural speech to train the natural speech model $\lambda_{natural}$. The testing data consist of 1,500 sessions of natural speech (original waveform from NIST 2006 SRE corpus), 1,000 sessions of GMM-based converted speech and 1,000 sessions of unit-selection based converted speech. The duration of each sessions of natural speech or converted speech is roughly 5 minutes.

4.2. GMM-based converted speech are available

Assume that we have 100 sessions of GMM-based converted speech for training of $\lambda_{converted}$. The performance of the three different features are presented in Table 1. With cos-phase and MGDF-phase features, the EER is reduced from 16.8% of MFCCs to 6.60% and 9.13%, respectively.

Table 1: *Equal error rate (EER, %) of detection performance when GMM-based converted data is available for training.*

Feature	Equal Error Rate (%)
MFCCs	16.80
cos-phase	6.60
MGDF-phase	9.13

4.3. Unit-selection based converted speech are available

Assume that we have 100 sessions unit-selection based converted speech to estimate the parameters of the converted speech model $\lambda_{converted}$. The detection results are presented in Table 2. We note that the cos-phase and MGDF-phase features reduce the EER from 15.35% of MFCCs to 3.93% and 4.60%, respectively.

Table 2: *Equal error rate (EER, %) of detection performance when unit-selection based converted data is available for training.*

Feature	Equal Error Rate (%)
MFCCs	15.35
cos-phase	3.93
MGDF-phase	4.60

4.4. No converted speech data are available

In this case, we assume that neither GMM-based nor unit-selection based converted speech is available for training, but the speech analysis module and synthesis filter are available. Therefore, we extract parameters from the 100 sessions of natural speech for training natural speech model $\lambda_{natural}$, and then pass the parameters directly to the synthesis filter to reconstruct the waveforms. We then use the 100 resulting sessions to train the converted speech model $\lambda_{converted}$. The detection error tradeoff (DET) curves are presented in Figure 2. We can see

that even without converted speech for training the detector, the performance of the features derived from phase spectrum outperform the MFCCs tremendously. The cos-phase and MGDF-phase features reduce the EER from 20.20% of MFCCs to 5.95% and 2.35%, respectively.

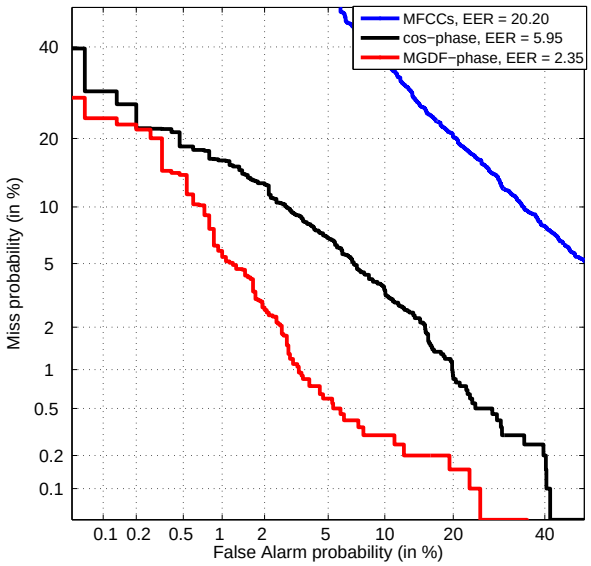

Figure 2: *DET curve of detection performance when converted data is NOT available for training.*

5. Conclusions

In this study, the features, cos-phase and MGDF-phase, derived from phase spectrum are adopted to detect converted speech for spoofing attack to enhance the security of speaker verification system. Under three different situations, the proposed features outperform the MFCCs consistently, especially when converted speech is not available for training. To detect converted speech in the paradigm of Figure 1, we have shown that our analysis-synthesis method offers an effective alternative solution that simulates converted data when actual voice transformation techniques are unavailable.

6. References

[1] J.P. Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D.A. Reynolds, "An overview of automatic speaker recognition technology," in *ICASSP 2002*.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, March 1998.

[4] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP 1998*.

[5] Y. Stylianou, "Voice transformation: a survey," in *ICASSP 2009*.

[6] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *EUROSPEECH 1999*.

[7] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *ICSLP 2000*.

[8] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using a HMM-based speech synthesis system," in *EUROSPEECH 2001*.

[9] P. DeLeon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Odyssey 2010*.

[10] Qin Jin, Arthur Toth, Alan W. Black, and Tanja Schultz, "Is voice transformation a threat to speaker identification?," in *ICASSP 2008*.

[11] Q. Jin, A.R. Toth, T. Schultz, and A.W. Black, "Voice convergin: Speaker de-identification by voice transformation," in *ICASSP 2007*.

[12] J.F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Interspeech*, 2007.

[13] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech," in *ICASSP 2012*.

[14] Z. Wu, E. S. Chng, and H. Li, "Speaker verification system against two different voice conversion techniques in spoofing attacks," Technical report: available at http://www3.ntu.edu.sg/home/wuzz/.

[15] P.L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *ICASSP 2011*.

[16] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics letters*, vol. 45, no. 7, pp. 381–383, 2009.

[17] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[18] R.M. Hegde, H.A. Murthy, and V.R.R. Gadde, "Significance of the modified group delay feature in speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 190–202, 2007.

[19] K.K. Paliwal and L.D. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[20] L.D. Alsteris and K.K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[21] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP 1992*.

[22] "Speech Signal Processing Toolkit (SPTK) version 3.4," *Software available at http://sp-tk.sourceforge.net/*.