A Study of the Vulnerability of Text-Dependent Speaker Verification System Against Voice Conversion Spoofing Attack

Zhizheng Wu, Anthony Larcher, Kong Aik Lee, Eng Siong Chng, Tomi Kinnunen, Haizhou Li

A Study of the Vulnerability of Text-Dependent Speaker Verification System Against Voice Conversion Spoofing Attack

Zhizheng Wu[1,2]*, Anthony Larcher*[3]*, Kong Aik Lee*[3]*, Eng Siong Chng*[1]*, Tomi Kinnunen*[4]*, Haizhou Li*[1,3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3]Human Language Technology department, Institute for Infocomm Research, Singapore
[4]School of Computing, University of Eastern Finland, Finland

wuzz@ntu.edu.sg

Abstract

Voice conversion technique, which is to change one speaker's voice to sound like it was pronounced by another speaker, has the potential to break down a speaker verification system. The vulnerability of text-independent speaker verification systems under spoofing attack simulated by statistical voice conversion has been confirmed in our previous work. In this study, we continue the study of vulnerability of the text-dependent speaker verification system attacked voice conversion techniques. We implemented both the joint density Gaussian mixture model (JD-GMM) based voice conversion and the unit-selection based voice conversion systems to simulate spoofing attack. In addition, the performances of text-independent and text-dependent speaker verification systems are compared. We conduct the experiments using RSR2015 database which is recorded using mobile device. The experiments show that .....

Index Terms: Speaker verification, text-dependent, text-independent, voice conversion, spoofing attack, security

1. Introduction

A large number of measurements have been investigated for biometric recognition systems. One of the most popular measurements is voice, which is easy to collect and use. To automatically and accurately verify the claimed identity of a speaker based on the speaker's speech sample is the main task of speaker verification. There are two kinds of speaker verification systems: text-independent speaker verification (TD-SV) and text-dependent speaker verification (TI-SV). TD-SV requires the speaker to speak a specific textual transcription, while TI-SV does not have this constraint and allows the speaker to speak anything for verification. Therefore, TD-SV is able to make use of phonetic/linguistic information to make the decision. Both TD-SV and TI-SV have many applications in access control systems, such telephone banking [1], to protect personal secret and privacy. Thus, the security of such verification system is the major concern to the clients.

To respond to such concern, the vulnerability of speaker verification systems under spoofing attacks has been studied. Several methods have been employed to simulate the spoofing attack, including replay attack [2, 3], human voice mimicking [4] and artificial signal spoofing[5]. The above spoofing techniques are not so flexible to generate the claimed speaker's voice, especial to generate a voice uttering a specific transcription which is required in a text-dependent speaker verification system. Due the popular available of speech synthesis and voice conversion techniques, speech synthesis and voice conversion methods become to be the easiest available techniques for the attackers, and these techniques are seriously threat to speaker verification system. In [6], the authors use an adapted HMM-based speech synthesis system, which is flexible to generate one speaker's voice given the transcripts, to simulate the spoofing attack. In [7], voice conversion technique is employed to simulate the spoofing attack, and text-independent speaker verification systems with and without high level text-constraint information are compared. In addition to the studies using high quality speech, spoofing attack studies are also carried out using telephone quality speech. In [8, 9], voice conversion technique is adopted to convert telephone quality speech to attack several different speaker verification systems including the classic GMM-UBM system and the state-of-the-art joint factor analysis system .

However, above spoofing attack studies are all employing text-independent speaker verification systems, which do not utilize phonetic information. It has been shown that the performance of the text-independent speaker verification systems can be degraded to an unacceptable level [8, 9]. However, whether a text-dependent speaker verification system is robust against spoofing attack is still an open question. Due to the popular of smart phones or mobile devices, the speaker verification technique starts to have application in unlocking the phone or devices [10]. To this end, the security of speaker verification systems is important to mobile devices users. To respond to this question, in this study, we focus on evaluating the performance of speaker verification systems under spoofing attach on smart phone or mobile device. We adopt both text-dependent speaker verification system and text-independent speaker verification for comparison. In addition, we use two voice conversion methods: joint density Gaussian mixture model and unit selection methods, to simulate the spoofing attack.

2. Voice conversion techniques

The task of voice conversion is to modify one speaker's (source) voice to sound like it was uttered by anther speaker (target) while keeping the linguistic information. Thus, it has potential ability to break down both text-dependent and text-independent speaker verification systems. In a typical voice conversion system, it consists of off-line training and run-time conversion processes. During off-line training, a relationship between the source and target speech is established. In the run-time conversion process, the relationship is applied to the input testing speech to generate the converted speech signal. In this study, we

employ two voice conversion systems to simulate the spoofing attack.

2.1. GMM-based voice conversion

The first voice conversion method is based on the joint density Gaussian mixture model (JD-GMM), which is originally proposed in [11] and become the mainstream approach [12, 13].

Given a parallel training data from source $\mathbf{X}$ speaker and target $\mathbf{Y}$ speaker, dynamic time warping is employed to

The training data of source speech contains N frames of spectral vectors $\mathbf{X} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top, \ldots, \mathbf{x}_N^\top]^\top$, where $\mathbf{x}_n \in \mathcal{R}^d$, and the training data of target speech contains M frames of spectral vectors $\mathbf{Y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_m^\top, \ldots, \mathbf{y}_M^\top]^\top$, where $\mathbf{y}_m \in \mathcal{R}^d$. For parallel data, we can use dynamic time warping (DTW) algorithm to align source feature vectors to their counterparts in the target; for non-parallel data, non-parallel frame alignment method used in [**?**, **?**] can be adopted to obtain feature vector pairs $\mathbf{Z} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \ldots, \mathbf{z}_t^\top, \ldots, \mathbf{z}_T^\top]^\top$, where $\mathbf{z}_t^\top = [\mathbf{x}_n^\top, \mathbf{y}_m^\top]^\top \in \mathcal{R}^{2d}$.

The joint probability density of X and Y is modeled by GMM as in (1):

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{Z}) = \sum_{l=1}^{L} w_l^{(z)} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_l^{(z)}, \boldsymbol{\Sigma}_l^{(z)}) \qquad (1)$$

where $\boldsymbol{\mu}_l^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_l^{(x)} \\ \boldsymbol{\mu}_l^{(y)} \end{bmatrix}$ and $\boldsymbol{\Sigma}_l^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_l^{(xx)} & \boldsymbol{\Sigma}_l^{(xy)} \\ \boldsymbol{\Sigma}_l^{(yx)} & \boldsymbol{\Sigma}_l^{(yy)} \end{bmatrix}$ are the mean vector and the covariance matrix of the multivariate Gaussian density $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_l^{(z)}, \boldsymbol{\Sigma}_l^{(z)})$, respectively. Given the component l, $w_l^{(z)}$ is its prior probabilities with $\sum_{l=1}^{L} w_l^{(z)} = 1$.

In the training phase, the GMM parameters $\lambda^{(z)} = \{w_l^{(z)}, \mu_l^{(z)}, \boldsymbol{\Sigma}_l^{(z)} | l = 1, 2, \ldots, L\}$ are estimated using the expectation maximization (EM) algorithm in maximum likelihood sense.

In the conversion phase, given a source speech feature vector $\mathbf{x}$, the joint density model is adopted to formulate a transformation function to predict the target speaker's feature vector $\hat{\mathbf{y}} = F(\mathbf{x})$, as follows:

$$\begin{aligned} F(\mathbf{x}) &= E(\mathbf{y}|\mathbf{x}) \\ &= \sum_{l=1}^{L} p_l(\mathbf{x})(\boldsymbol{\mu}_l^{(y)} + \boldsymbol{\Sigma}_l^{(yx)}(\boldsymbol{\Sigma}_l^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_l^{(x)})), \end{aligned}$$

$$p_l(\mathbf{x}) = \frac{w_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l^{xx})}{\sum_{k=1}^{L} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})}$$

where $p_l(\mathbf{x})$ is the posterior probability of source vector $\mathbf{x}$ belonging to the l^{th} Gaussian component.

The transformation function is applied to the source speech feature vectors, then the converted feature vectors are passed to speech synthesis vocoder to reconstruct audible speech signals.

2.2. Unit-selection based voice conversion

Different from GMM-based voice conversion, unit-selection based method directly makes use of the original training data, instead of transforming the source speech to the target speech space. In this section, we will introduce implementation the unit-selection based voice conversion method used in this study.

Similar as that in the GMM-based voice conversion method, dynamic time warping is employed to find the frame alignment $\mathbf{Z}$ of the parallel data from source X and target Y speakers. During conversion, each input source speech vector $\hat{\mathbf{x}}_\tau$ is paired up with a source speech vector $\mathbf{x}_t$ from training data. Thus, the aligned target vector $\mathbf{y}_t$ is used the converted speech vector for $\hat{\mathbf{x}}_\tau$.

2.3. Setups of voice conversion

3. Speaker verification systems

In this study, we compare the performance of text-dependent and text-independent speaker verification systems under spoofing attack. In this section, we introduce the setups of the two speaker verification systems.

3.1. Text-Dependent Speaker Verification System

3.2. Text-Independent speaker verification system

4. Database

In this study, we employ the RSR2015 database [14] to design the spoofing attack database and conduct the speaker verification experiments. The RSR2015 database has three parts. In part 1, each speaker reads 30 utterances and the average duration of each utterance is 3.2 seconds; in part 2, each speaker is asked to pronounce short command; while in part 3, randomly prompted digit sequences is recorded by each speaker. 300 speakers including 157 male and 143 female speakers take part in the whole recording process. In speaker verification task, we use part 1 for speaker verification experiments. We note that the verification test is gender dependent and only the trials with match transcripts are used in the test. The trials statistics of the speaker verification test is presented in Table 4.

Table 1: Statistics of the trials in the speaker verification test

	Development		Evaluation	
	Male	Female	Male	Female
Target speakers	50	47	57	49
Impostor trials	437,631	389,160	573,664	423,312
Genuine trials	xx	xx	xx	xx

To generate the spoofing database, we use part 2 of the RSR2015 database, which has different transcripts from part 1, to train the conversion functions or to find frame pairs between source and target speech. We use the conversion functions or frame pairs to convert the impostor's speech to the target speaker, making the converted speech sound like it was uttered by the target speaker. We note that only the impostor trials in the speaker verification data set are converted while the genuine trials are kept as the same as original.

5. Experimental results and discussion

6. Conclusions

7. References

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

[2] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proceedings of the European Conference on Speech Communication and Technology*, vol. 3, 1999, pp. 1211–1214.

[3] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 10 workshop*, 2010, pp. 131–134.

[4] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "How vulnerable are prosodic features to professional imitators?" in *The Speaker and Language Recognition Workshop (Odyssey 2008)*, 2008.

[5] F. Alegre, R. Vipperla, N. Evans *et al.*, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, 2012.

[6] P. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, 2012.

[7] Q. Jin, A. Toth, A. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?" in *ICASSP 2008*. IEEE, 2008, pp. 4845–4848.

[8] Z. Wu, T. Kinnunen, E. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *APSIPA ASC*, 2012.

[9] T. Kinnunen, Z. Wu, K. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4401–4404.

[10] "Lenovo A586 touts voice unlock through Baidu, A*STAR verification tech," *http://www.engadget.com/2012/12/01/lenovo-a586-touts-voice-unlock-through-baidu-astar/*.

[11] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 285–288.

[12] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[13] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1784–1794, 2012.

[14] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The rsr2015: Database for text-dependent speaker verification using multiple pass-phrases."