Speaker verification system against two different voice conversion techniques in spoofing attacks

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

[1]School of Computer Engineering, Nanyang Technological University (NTU), Singapore
[2]Human Language Technology Department, Institute for Infocomm Research (I^{2}R), Singapore

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

Abstract

Voice conversion technique, which modifies one's (source speaker) voice to sound like another (target speaker), is a threat to automatic speaker verification. In this paper, we present new results evaluating the current state-of-the-art speaker verification system, Gaussian mixture model supervector with joint factor analysis (GMM-JFA) system, against spoofing attacks. The spoofing attacks are simulated by two voice conversion techniques: Gaussian mixture model based conversion method and unit selection based conversion method. A subset of the core task, 1conv4w-1conv4w, in the National Institute of Standards and Technology (NIST) 2006 speaker recognition evaluation (SRE) corpus is used to evaluate the performance of spoofing attacks. The results show that GMM-based conversion method which increases the false acceptance rate (FAR) from 3.24% to 17.33%, while the unit selection based method increases the FAR from 3.24% to 32.54%. This suggests that GMM-JFA system is more vulnerable towards unit selection based conversion than GMM-based conversion.

1. Introduction

The objective of speaker verification is to make a binary decision to accept or reject a claim of identity based on the user's speech samples [1, 2]. In practice, speaker verification system can be used to verify a speaker's identity for control access to services such as telephone banking [1], voice mail [1, 2], and so on. On the other hand, the task of voice conversion is to modify one speaker's voice (source speaker) so that is sounds as if it has been uttered by another speaker (target speaker) [3, 4]. This paper studies voice conversion and speaker verification in an attack and defence experiment. We assume that the converted speech samples are obtained in telephony conversations where voice conversion is performed in one of the speakers.

There have been multiple studies in voice conversion vs speaker verification. For example, speaker verification system against imposter's speech which are generated from HMM-based speech synthesis system [5] or adapted speech synthesis system with small size adaptation data [6], and voice conversion techniques [7, 8, 9]. These studies are all carried out on high quality speech. In telephone applications, such as telephone banking, the speaker verification system has to deal with telephone speech which is low quality and affected by channel variability. In our previous research, we conducted spoofing attack study using telephone speech [10] with five speaker verification systems: GMM-UBM (Gaussian mixture model with universal background model) system [11], VQ-UBM (vector quantized codebook with universal background model) system [12], GLDS-SVM (generalized linear discriminant sequence kernel support vector machine) system [13], GMM-SVM system [14], and GMM-JFA (Gaussian mixture model supervector with joint factor analysis) system [10]. Our previous results suggested that the GMM-JFA system, which is the current state-of-the-art speaker verification system, obtained the best performance against spoofing attack simulated by a simple voice conversion technique.

In this study, we continue the study of vulnerability of the current state-of-the-art speaker verification system by examining the performance of GMM-JFA system against spoofing attack. We will use two different voice conversion methods, GMM-based conversion method and unit selection based method, to simulate the spoofing attack. In the previous study, we used GMM-based voice conversion method to simulate spoofing attack. In the GMM-based voice conversion method, the transformation parameters is derived from Gaussian mixture models (GMM), and then the linear transformation is applied to the spectrum parameters of the source speech frames. Although GMM-based voice conversion techniques can generate speech with acceptable quality, the transformation is not perfect, and hence may not transform the source feature vector to the target feature vector space. That is the reason why informal listening tests show that the converted speech may not resemble the target speaker, and the converted speech may sound like another speaker who is neither source speaker nor target speaker. On the other hand, for telephone speech conversion, GMM-based conversion method can be viewed as a joint shift of channel factor and speaker characteristic. While in the unit-selection based conversion method, target speaker's feature vectors are directly used to synthesize the converted speech, without changing the original spectral envelop. If we consider the resulting speech a collection of speech frame regardless of the continuity and prosody of speech flow, unit selection should produce speech that sounds closer to the target speaker.

Although informal listening tests show that converted speech from GMM-based conversion method is much smoother than that from unit-selection based method, current speaker verfication systems, such as GMM-JFA system, are not considering the naturalness of the speech. One can expect that GMM-JFA speaker verification system is more vulnerable to unit selection based voice conversion.

This paper is organized as follows. In sections 2, we will describe voice conversion techniques based on Gaussian mixture model and unit selection; the speaker verification system used in this study will be presented in section 3. In section 4, the experimental setups and results are presented and discussed. We conclude this article in section 5.

2. Voice conversion methods

We study two different voice conversion techniques in simulating the spoofing attacks. One is GMM-based conversion, which trains a mapping function between source and target, and requires a parallel corpus for training. The other one is unit selection based voice conversion technique, which doesn't require training data from the source speaker. We will introduce the two conversion methods briefly in this section.

2.1. GMM-based voice conversion

The most popular voice conversion method is based on joint density Gaussian mixture model (GMM), which is originally proposed in [4]. We apply this method to simulate spoofing attack in this study and describe as follows.

The training data of source speech contains N frames spectral vectors $\mathbf{X} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top, \ldots, \mathbf{x}_N^\top]^\top$, and the training data of target speech contains M frames spectral vectors $\mathbf{Y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_m^\top, \ldots, \mathbf{y}_M^\top]^\top$. For parallel data, we can use dynamic time warping algorithm to align source feature vectors to their counterparts in the target; for non-parallel data, non-parallel frame alignment method used in [15, 10] 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 $\mathbf{z}_t^\top = [\mathbf{x}_n^\top, \mathbf{y}_m^\top]^\top$.

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 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 the prior probabilities of z, and $\sum_{l=1}^{L} w_l^{(z)} = 1$

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.

While 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 $\hat{\mathbf{y}} = F(\mathbf{x})$, the transformation function $F(.)$ is given 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 an audible speech signal.

2.2. Unit selection based voice conversion

In the GMM-based voice conversion method, both source and target speech are required to estimate a transformation function. For conversion of telephone speech, the transformation can be viewed as a joint shift of the speaker characteristic and the channel factor. As unit selection method directly uses target speaker's voice to synthesize new speech, the resulting speech frames match well the voice of the target speaker.

In this study, we follow the unit selection approach proposed in [16]. However, we use a different cost function, as in this study, we are not so interested in the continuity of the speech. The unit selection based voice conversion method is described as follows.

Given target speech, we first extract feature vector from the speech signal and obtain N frames target feature vectors $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N]$.

Then, give a feature vectors sequence from source speaker, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$, we must determine a vector sequence from target feature vector space to best fit the given source vector sequence. So every frame $\mathbf{x}_t$ from source vector sequence is paired up with the nearest feature vector $\mathbf{y}_n^{'}$ from target feature vector space in Euclidean distance sense.

After that, the paired feature vectors from the target space are concatenated and passed to speech synthesis vocoder to reconstruct an audible speech signal.

3. Speaker verification system

In this study, we only considered the GMM-JFA system. As in our previous study [10], we consider five speaker verification systems, GMM-UBM [11, 17], VQ-UBM [12], GLDS-SVM [13], GMM-SVM [14] and GMM-JFA [18], and GMM-JFA system is the best system against spoofing attacks [10].

The GMM-JFA system, which adopts joint factor analysis technique for modeling intersession and speaker variability in the GMM supervector space, is a widely recognized high performance system [18]. In this study, the GMM-JFA system uses 512 Gaussian mixtures. The Gaussian mixture model is trained using the HTK toolkit [19]. For score normalization, we use T-norm followed by Z-norm (TZ-norm). The National Institute of Standards and Technology (NIST) 2004 speaker recognition evaluation(SRE), NIST 2005 SRE, MIXER 5 and Switchboard corpora are used to train eigenchannel, eigenvoice and diagonal models, as well as the T-norm and Z-norm cohort models.

We also use the same acoustic front-end to extract acoustic feature as previous study [10]. 12 dimensions MFCCs with delta and delta-delta coefficients are computed via 27-channel mel-frequency filterbank. Then RASTA filtering, voice activity detection and utterance level cepstrum mean variance normalization techniques are applied to the extracted MFCCs. So the final feature vectors are 36-dimension MFCCs.

4. Experimental Setup

4.1. Corpus

In this study, we use a subset of the core task, 1conv4w-1conv4w, in the National Institute of Standards and Technology (NIST) 2006 speaker recognition evaluation (SRE) corpus. In this subset task, there are 298 female and 206 male unique speakers and these speakers make 6,760 gender matched verification trials, including 3,978 genuine and 2,782 imposter trials. We try to make the gender and imposter-genuine pairs as balance as the original NIST 2006 SRE evaluation. Details of the

new corpus are shown in Table 1

Table 1: Statistics of the new trials (subset of NIST 2006 SRE core task).

	Female	Male	Total
Unique speakers	298	206	504
Genuine trials	2,349	1,629	3,978
Impostor trials	1,636	1,146	2,782

We then design the spoofing corpus. Suppose we have the key for each trial, in other words, we know whether a trial is an impostor or a genuine test.. We keep the speaker models the same as that in the baseline speaker verification test, but process the test utterances which are impostors through voice conversion system. Hence, the 3,978 genuine trials are kept as original, while the 2,782 impostor trials are processed through corresponding conversion functions, which have been trained in advance for different speaker pairs. To keep the utterances used for training the speaker enrollment models and that for training voice conversion functions disjoint, we make use of the *3conv4w* and *8conv4w* training sections in the NIST 2006 SRE corpus in the voice conversion part.

For voice conversion, the feature extraction and waveform reconstruction are done as follows: The sampling rate of the speech files is 8,000 Hz. The speech signal is windowed using 25 ms Hamming window with a 5ms shift. 30 dimension mel-cepstral coefficients ,which are used to represent spectrum, are extracted using the Speech Signal Processing Toolkit (SPTK) tool [20]. Only voiced frame are passed to the voice conversion system. Fundamental frequency (F0) values are automatically extracted using the robust algorithm for pitch tracking (RAPT) algorithm [21]. F0 conversion is done by equalizing the means and variances of the source and target log-F0 distributions. After both spectral and F0 conversion are done , SPTK tool is also used to reconstruct waveform.

4.2. Results and analysis

To evaluate the spoofing attack performance, we report the equal error rate (EER) and the minimum value of the detection cost function (MinDCF). The larger of EER and MinDCF, the better spoofing attack of the conversion.

Table 2: Performance of GMM-JFA under spoofing attack using two different voice conversion techniques.

Voice conversion	Equal error rates	100 × MinDCF
Baseline (*No conversion*)	3.24%	1.57
GMM-based	7.61%	3.49
unit-selection based	11.58%	5.98

The EER and MinDCF results of GMM-JFA system under two different spoofing corpora are shown in Tabel 2. We can see that, comparing with GMM-based voice conversion method, unit selection based method can significantly increase the EER and MinDCF. The reason is that GMM-based method tries to transform the source speech to the target speech, while unit selection based method directly use target speech to synthesize new speech to simulate spoofing attack. In the telephone speech case, GMM-based method transforms both the channel factor and speaker characteristic to the target channel factor and speaker characteristic. While the unit selection based method does not change the target speaker's characteristic (except for the prosody, mainly duration), and use a different channel factor, for which GMM-JFA system is more robust.

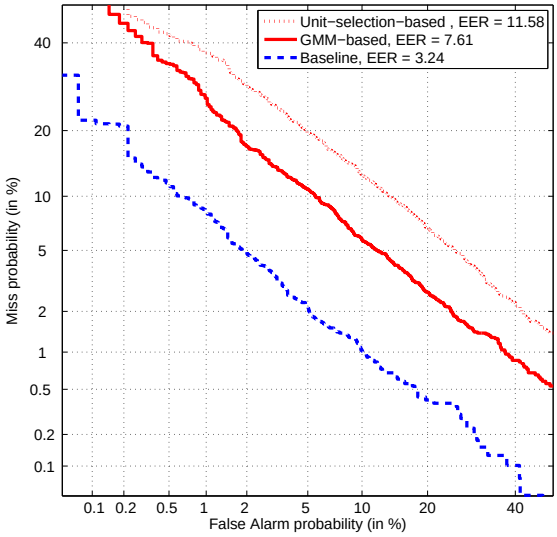

Figure 1: *DET for baseline and two converted impostor attack on NIST 2006 SRE.*

In real application, the decision thresholds for EER and MinDCF are made on original data, as the speaker verification system would not know whether there is spoofing attack or not. Hence, we set the decision thresholds on the original baseline data, and then apply these fixed thresholds to the converted data. The false acceptance rates on the spoofing data using the two different conversion are reported in Table 3.

Table 3: False acceptance rates (FAR, %) for spoofing attack when decision threshold is set to EER point on the baseline data

Voice conversion	FAR (%)
None (Baseline)	3.24
GMM	17.33
unit selection	32.54

5. Conclusion

In this study, we investigate the effect of artificially modified speech using two different voice conversion techniques on the GMM-JFA speaker verification system. Without surprise, voice conversion based on unit selection presents a greater threat to the existing speaker verification system. As voice conversion based unit selection directly use target speaker's voice to synthesize new speech, in the telephone speech case, unit selection uses the target speaker's speech and a new channel factor to synthesize new speech, while GMM-based voice conversion techniques jointly shift speaker characteristic and channel factor of source speech to target; and current speaker verification system does not consider the naturalness of the speech.

As a human listener can easily discriminate natural speech

and synthesized/converted speech, for which, current state-of-the-art speaker verification system can not do this, in future work, we will develop techniques to detect natural speech and synthesized/nonatural speech for speaker verification system.

6. Acknowledgement

The authors would like to thank Dr. Lee Kong Aik from Institute for Infocomm Research (I^2R), Singapore for providing GMM-JFA system and related databases for training; and Dr. Tomi Kinnunen from University of Eastern Finland, Joensuu, Finland for providing acoustic front-end to extract feature for speaker verification system.

7. 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 *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 4, pp. IV–4072.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, 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 *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*. IEEE, 1998, vol. 1, pp. 285–288.

[5] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using a HMM-based speech synthesis system," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, September 2001, pp. 759–762.

[6] P. DeLeon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 151–158 (paper 28).

[7] J.F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[8] Q. Jin, A.R. Toth, A.W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4845–4848.

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

[10] 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 *Acoustics, Speech and Signal Processing, 2012. Proceedings of the 2012 IEEE International Conference on*. IEEE, 2012.

[11] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[12] V. Hautamki, T. Kinnunen, I. Krkkinen, M. Tuononen, J. Saastamoinen, and P. Frnti, "Maximum *a Posteriori* estimation of the centroid model for speaker verification," *IEEE Signal Processing Letters*, vol. 15, pp. 162–165, 2008.

[13] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[14] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[15] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 944–953, 2010.

[16] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[17] T. Kinnunen, J. Saastamoinen, V. Hautamki, M. Vinni, and P. Frnti, "Comparative evaluation of maximum *a Posteriori* vector quantization and Gaussian mixture models in speaker verification," *Pattern Recognition Letters*, vol. 30, no. 4, pp. 341–347, March 2009.

[18] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," technical report CRIM-06/08-14, 2006.

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The htk book (for htk version 3.4)," 2006.

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

[21] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.