A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case

Zhizheng Wu*†‡, Tomi Kinnunen‡, Eng Siong Chng*, Haizhou Li*§¶, Eliathamby Ambikairajah¶
* School of Computer Engineering, Nanyang Technological University, Singapore
† Temasek Laboratories@NTU, Nanyang Technological University, Singapore
‡ School of Computing, University of Eastern Finland, Joensuu, Finland
§ Human Language Technology Department, Institute for Infocomm Research , Singapore
¶ School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia
wuzz@ntu.edu.sg, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—Voice conversion technique, which modifies one speaker's (source) voice to sound like another speaker (target), presents a threat to automatic speaker verification. In this paper, we first present new results of evaluating the vulnerability of current state-of-the-art speaker verification systems: Gaussian mixture model with joint factor analysis (GMM-JFA) and probabilistic linear discriminant analysis (PLDA) systems, against spoofing attacks. The spoofing attacks are simulated by two voice conversion techniques: Gaussian mixture model based conversion and unit selection based conversion. To reduce false acceptance rate caused by spoofing attack, we propose a general anti-spoofing attack framework for the speaker verification systems, where a converted speech detector is adopted as a post-processing module for the speaker verification system's acceptance decision. The detector decides whether the accepted claim is human speech or converted speech.**

A subset of the core task in the NIST SRE 2006 corpus is used to evaluate the vulnerability of speaker verification system and the performance of converted speech detector. The results indicate that both conversion techniques can increase the false acceptance rate of GMM-JFA and PLDA system, while the converted speech detector can reduce the false acceptance rate from 31.54% and 41.25% to 1.64% and 1.71% for GMM-JFA and PLDA system on unit-selection based converted speech, respectively.

I. Introduction

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 controlling access to services such as telephone banking [1] or voice mail [1], [2]. In a related speech technology, the voice conversion research, whose task is to modify one speaker's (source) voice so that is sounds as if it has been uttered by another speaker (target) [3], [4], can be potentially used to attack or fool speaker verification systems.

There have been a number of studies on speaker verification system security against impostor attacks. For example, playback attacks [5], [6], HMM-based speech synthesis system [7], speaker-adapted speech synthesis [8], [7], [9], voice conversion [10], [11], [12] and even human voice mimicking [13], [14]. These studies have clearly shown that false acceptance rate of speaker verification systems increases under various spoofing attacks.

While the above studies are all carried out on high quality speech, in telephone applications, such as telephone banking, speaker verification system has to cope with low-quality signal. In addition, the signal is affected by channel variability and is also distorted during transmission. These factors may affect the speaker characteristic in the signal. In our previous study, we also conducted spoofing attack study using telephone speech [15] with five speaker verification systems: GMM-UBM (Gaussian mixture model with universal background model) [16], VQ-UBM (vector quantized codebook with universal background model) [17], GLDS-SVM (generalized linear discriminant sequence kernel support vector machine) system [18], GMM-SVM (Gaussian mixture model with support vector machine) [19], and GMM-JFA (Gaussian mixture model supervector with joint factor analysis) [15]. These studies on telephone speech also indicate vulnerability of the current speaker verification systems under spoofing attack.

While there are reports on synthetic or converted speech overcoming state-of-the-art speaker verification system, listening tests have shown that human can easily distinguish natural speech from synthetic or converted speech. To enhance the security of speaker verification system, a detector for distinguishing converted speech and natural speech will be necessary. In [20], the authors proposed to make use of the differences in the relative phase shift between high quality human and synthetic speech to differentiate synthetic from natural speech. In that study, synthetic speech is assume to be available for training the detector. In our previous work [21], we proposed to use phase spectrum features to detect synthetic speech. Two features were derived: cosine normalized phase and modified group delay phase. We tested the features under three different assumptions of the training data for the converted speech detector: a) only GMM based converted speech is available; b) only unit selection based converted speech is available; c) no converted speech is available. When no converted speech is available, we assume that the vocoder used in the analysis and synthesis stages of the conversion system is available and use the analysis-synthesis speech to train the converted speech model.

In this study, we look into spoofing attacks simulated by two conversion techniques: GMM based conversion [3], [4]

and unit-selection based conversion [22]. The two methods are the popular voice conversion methods. To reduce the false acceptance rate caused by spoofing attack, we then integrate the converted speech detector with the speaker verification system for anti-spoofing. The performance of GMM-JFA [23] and probabilistic linear discriminant analysis (PLDA) systems [24] are evaluated.

II. Speaker verification systems

In this study, we consider two state-of-the-art speaker verification systems: GMM-JFA and PLDA systems. The two systems use the same acoustic front-end to extract acoustic feature[15]. 12 dimensions MFCCs with delta and delta-delta coefficients are computed via 27-channel mel-frequency filterbank. Then RASTA filtering, energy-based 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.

A. Joint factor analysis system

The GMM joint factor analysis (GMM-JFA) models intersession and speaker variability in the GMM supervector space [23]. GMM-JFA system decomposes a supervector for a speaker into speaker-independent, speaker-dependent, channel-dependent and residual components where each component is represent by a low-dimensional set of factors via factor loadings. In this study, the GMM-JFA system uses 512 Gaussian mixtures. The Gaussian mixture model is trained using the HTK toolkit [25]. For score normalization, we use T-norm followed by Z-norm (TZ-norm). NIST SRE 2004, NIST SRE 2005, MIXER 5 and Switchboard corpora are used to train the eigenchannel, eigenvoice and diagonal models, as well as the T-norm and Z-norm cohort models.

B. Probabilistic linear discriminant analysis system

Probabilistic linear discriminant analysis (PLDA) is a generative model which is similar to the JFA approach. Rather than using GMM supervectors as the basis for factor modeling, PLDA models *i-vector* distributions for speaker verification [24], [26], and model session and speaker variability within the i-vector space. The i-vector is a low-dimensional set of factors, and it is used to represent speaker and channel variability via factor loadings. The PLDA system also use 512 Gaussian mixtures. NIST 2004 SRE, NIST 2005 SRE and Switchboard corpora are used to estimate the total variability matrix (factor loadings). An i-vector is extracted from each session of these corpora for training the PLDA model. To deal with the non-Gaussian behavior of the i-vectors, length normalization is performed prior to train PLDA model [27]. The dimensionality of the i-vector is empirically set to be 400.

III. Voice conversion methods

We study two different voice conversion techniques to simulate 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 technique is unit selection based voice conversion technique, which does not require training data from the source speaker. We will introduce the two conversion methods briefly in this section.

A. 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].

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 [28], [15] 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:

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

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.

B. 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. Since unit selection method directly uses the 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 unit selection approach similar to [22], while we use a simplified cost function for easy implementation. 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, given feature vector 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. Every frame $\mathbf{x}_t$ from the source vector sequence is paired up with the nearest feature vector $\mathbf{y}_n^{'}$ from the 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.

In the GMM-based voice conversion method, the transformation function 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 the source speaker nor the target speaker. On the other hand, for telephone speech conversion, GMM-based conversion method can be viewed as a joint shift of the channel factor and the speaker characteristics. 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 envelops. If we consider the resulting speech as a collection of speech frames regardless of the continuity and prosody of speech flow, unit selection should produce speech that sounds closer to the target speaker. Hence, it is expected that speaker verification systems are more vulnerable to unit selection based conversion technique.

IV. Anti-spoofing attack in speaker verification

As converted speech can increase the false acceptance rate of the speaker verification system, it would be useful to incorporate a converted speech detector into a speaker verification system.

A. Modified group delay phase

In our previous work [21], we have shown that natural speech phase information is lost during the analysis-synthesis step for some vocoder. Hence, phase information could be used to detect converted speech. In this study, we use a modified group delay phase spectrum [29] to capture the fine structure of the group delay phase spectrum for converted speech detector.

Given a speech signal $x(n)$, the modified group delay (MGD) cepstral coefficients can be calculated as follows:

1) Compute the short-time Fourier transform (STFT) $X(w)$ and $Y(w)$ of $x(n)$ and $nx(n)$, respectively.
2) Compute the smoothed spectrum $S(w)$ of $|X(w)|$.
3) Compute the MGD phase spectrum $\tau_\gamma(\omega) = (X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega))/ \mid S(\omega) \mid^{2\gamma}$.
4) Reshape the MGD phase spectrum $\tau_{\alpha,\gamma}(\omega) = \mid \tau_\gamma(\omega) \mid^\alpha \tau_\gamma(\omega)/ \mid \tau_\gamma(\omega) \mid$.
5) Apply discrete cosine transform (DCT) to the MGD phase spectrum.
6) Keep 12 cepstral coefficients, excluding the 0th coefficient.

In this study, we set α and γ to 0.4 and 1.2, respectively, which are optimized in our previous work [21]. The delta and delta-delta coefficient of MGD cepstral coefficients are not used.

B. Converted speech detector for anti-spoofing

To reduce the false acceptance rate of the speaker verification system caused by spoofing attack, we incorporate an anti-spoofing mechanism by using a converted speech detector as shown in Fig. 1.

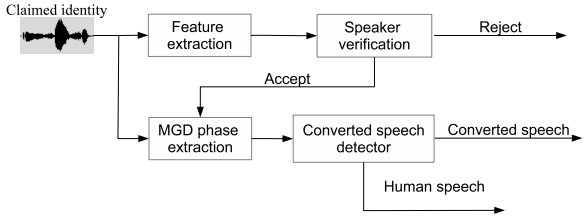

Fig. 1. A speaker verification system that incorporates an anti-spoofing mechansim

In the proposed speaker verification system, we extract MGD phase cepstral coefficients for converted speech detection. A claimed identity is accepted only if the speech is accepted by the speaker verification test and it is detected as human speech.

The converted speech detector is based on Gaussian mixture model, and we make the natural or converted speech decision based on the log likelihood ratio:

$$\Lambda(O) = \log p(O|\lambda_{\text{converted}}) - \log p(O|\lambda_{\text{natural}}) \qquad (2)$$

where O is the feature vector sequence of a speech signal, $\lambda_{\text{converted}}$ and λ_{natural} are GMM models for converted and natural speech, respectively. The number of Gaussian components of GMM is set to 512. The decision threshold is set to have an equal error rate on the development set [21].

V. Experiments

A. Data set

In this study, we use a subset of the verification trials in the core task, 1conv4w-1conv4w, of NIST SRE 2006. In this task,

there are 298 female and 206 male target speakers and 6,760 gender matched verification trials, including 3,978 genuine and 2,782 imposter trials, as summarized in Table I.

TABLE I
SUBSET OF NIST 2006 SRE CORE TASK IN THE EXPERIMENTS.

	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. As a result, the 3,978 genuine trials are kept as original, while the 2,782 impostor samples are processed to sound like those of the target speakers 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 SRE 2006 corpus for voice conversion function training.

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 [30]. 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) [31]. 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.

For synthetic speech detector, we randomly select 100 sessions of natural speech from training set for speaker verification system to train the natural speech model $\lambda_{\mathrm{natural}}$, and the analysis-synthesis data of these 100 sessions is used to train the converted speech model $\lambda_{\mathrm{converted}}$. The duration of each session is about 5 minutes conversational speech.

B. Results and analysis

1) Effect of spoofing attack: We first consider equal error rate (EER) and MinDCF (using the cost parameters in the SRE 2006 plan) of speaker verification systems under spoofing attack. The equal error rate (EER) and MinDCF values are presented in Table II. There are 3, 978 genuine trials for all conversion techniques, 2, 782 converted impostor trials for both GMM conversion and unit-selection conversion, and 2, 782 impostor trials for baseline which has no conversion. We find that both EER and MinDCF are increased because of spoofing attack. We also note that both GMM-JFA and PLDA are more vulnerable to unit-selection based conversion than GMM-based conversion.

TABLE II
PERFORMANCE OF GMM-JFA AND PLDA UNDER SPOOFING ATTACK USING TWO DIFFERENT VOICE CONVERSION TECHNIQUES.

Voice conversion	Equal error rates (EER %)		100 × MinDCF	
	GMM-JFA	PLDA	GMM-JFA	PLDA
Baseline (*No conversion*)	3.24	2.99	1.57	1.54
GMM conversion	7.61	6.77	3.49	2.76
Unit-selection conversion	11.58	11.18	5.98	3.89

As we only manipulate the imposter trials of the corpus, such a spoofing attack will only affect the false acceptance rate. In real world application, we typically set the decision threshold for EER and MinDCF on a development set of data. Let us use the original baseline data as the development set. So that we set the decision thresholds on the original baseline data, and then apply these fixed thresholds to the converted data. The false acceptance rates (FAR) on the spoofing data using the two different conversion are reported in Table III. We note that the FARs of the speaker verification systems are increased to unacceptable level under spoofing attacks.

TABLE III
EFFECT OF SPOOFING ATTACK ON FALSE ACCEPTANCE RATE (FAR %). SPEAKER VERIFICATION DECISION THRESHOLD IS SET TO EER POINT ON THE BASELINE CORPUS.

Voice conversion	GMM-JFA	PLDA
Baseline (*No conversion*)	3.24	2.99
GMM conversion	17.36	19.29
Unit-selection conversion	32.54	41.25

TABLE IV
EFFECT OF ANTI-SPOOFING ATTACK TO FALSE ACCEPTANCE RATE (FAR %). SPEAKER VERIFICATION DECISION THRESHOLD IS SET TO EER POINT ON THE BASELINE CORPUS.

Voice conversion	GMM-JFA	PLDA
Baseline (*No conversion*)	3.13	2.88
GMM conversion	0.0	0.0
Unit-selection conversion	1.64	1.71

2) Effect of anti-spoofing: Then, we evaluate the performance of proposed anti-spoofing attack framework which is presented in Fig. 1. The results are shown in Table IV, which are comparable with those in Table III. With the converted speech detector as a post-processing module for acceptance decision made by speaker verification system, the false acceptance rates are reduced to 0.0% for both GMM-JFA and PLDA under GMM conversion attack. And FARs are reduced from 32.54% and 41.25% to 1.64% and 1.71% of GMM-JFA and PLDA, respectively, under spoofing attack simulated by unit-selection conversion. While the performance of the baseline system without spoofing attack is not affected. We note that in Table IV, it does not mean that GMM conversion or unit-selection conversion gives smaller FARs than baseline which has no conversion, as there is no impostor trials in the GMM conversion or unit-selection conversion data sets, only has genuine trials and converted impostor trials.

VI. Conclusions

In this study, we first evaluated the vulnerability of two state-of-the-art speaker verification system to imposture using two voice conversion techniques. Experiments on telephone speech have indicated that voice conversion techniques could break down the state-of-the-art speaker verification systems: GMM-JFA and PLDA systems. The false acceptance rates of GMM-JFA and PLDA systems deteriorated to an unacceptable level for real application.

To manage the false acceptance rate and to enhance the security the speaker verification system, we proposed to integrate a converted speech detector with the speaker verification systems as presented in Fig. 1. The converted speech detector is based on phase spectrum feature, by assuming that the vocoder, which is used in analysis and synthesis of voice conversion system, is available. This may not be true in real situation. If a vocoder uses natural phase information in the synthesis step, our converted speech detector may not work. To enhance the converted speech detector, we will investigate vocoder independent features for training the converted speech detector in future.

VII. Acknowledgement

The authors would like to thank Dr. Kong-Aik Lee from Institute for Infocomm Research, Singapore for providing the GMM-JFA and PLDA speaker verification systems. The work of Tomi Kinnunen was supported by Academy of Finland (project 132129 and 253120).

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] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - a study of technical impostor techniques," in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, Budapest, Hungary, September 1999, pp. 1211–1214.

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

[7] 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.

[8] B.L. Pellom and J.H.L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," Phoenix, Arizona, USA, March 1999, pp. 837–840.

[9] 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).

[10] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech 2007 (ICSLP)*, Antwerp, Belgium, August 2007, pp. 2053–2056.

[11] Qin Jin, Arthur Toth, Alan W. Black, and Tanja Schultz, "Is voice transformation a threat to speaker identification?," Proc. ICASSP 2008, March 2008, pp. 4845–4848.

[12] 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.

[13] Y.W. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 2005, pp. 15–21.

[14] 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)*, Stellenbosch, South Africa, January 2008.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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*.

[21] Z. Wu, E. S. Chng, and H. Li, "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition," in *Interspeech*, 2012.

[22] 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.

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

[24] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, 2010.

[25] 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.

[26] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[27] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[28] 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.

[29] H.A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 1, pp. I–68.

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

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