Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints

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

[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, a technique to change one's voice to sound like that of another, poses a threat to even high performance speaker verification system. Vulnerability of text-independent speaker verification systems under spoofing attack, using statistical voice conversion technique, was evaluated and confirmed in our previous work. In this paper, we further extend the study to *text-dependent* speaker verification systems. In particular, we compare both joint density Gaussian mixture model (JD-GMM) and unit-selection (US) spoofing methods and, for the first time, the performances of text-independent and text-dependent speaker verification systems in a single study. We conduct the experiments using RSR2015 database which is recorded using multiple mobile devices. The experimental results indicate that text-dependent speaker verification system tolerates spoofing attacks better than the text-independent counterpart.

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

1. Introduction

Voice biometrics become popular especially in smartphone or telephony applications where voice services are provided. To automatically and accurately verify the claimed identity of a speaker based on just the speech sample is the main task of a speaker verification system [1]. Speaker verification encompasses two different input modes: *text-independent* speaker verification (TI-SV) and *text-dependent* speaker verification (TD-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 freely during enrolment and verification. TD-SV assumes cooperative speakers, while TI-SV doesn't. Both TD-SV and TI-SV are ideal for many access control applications, such as telephone banking or online transaction [2], whereby the objective is to protect personal secret and privacy. Thus, the reliability of such verification system is the major concern to the clients.

To address such concern, the vulnerability of speaker verification systems under spoofing attacks has been evaluated in many studies. Several methods have been employed to simulate the spoofing attack, including replay attacks [3, 4], human voice mimicking [5] and artificial signal spoofing [6]. Such methods do not really generate voice utterances of specific content required by a text-dependent speaker verification system.

The work of Tomi Kinnunen was supported by Academy of Finland (proj. no. 253120)

Speech synthesis and voice conversion techniques have become easily accessible for attackers, which pose a serious threat to the reliability of contemporary speaker verification system. In [7], the authors used an adapted HMM-based speech synthesis system, which is flexible to generate one speaker's voice given the transcripts, to simulate spoofing attacks. In [8], voice conversion technique was 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 [9, 10], voice conversion technique was 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.

In general, the above spoofing attack studies focus only on text-independent speaker verification systems, which do not directly utilize phonetic or linguistic information. Nevertheless, it has been shown that a text-independent speaker verification systems can be compromised to an unacceptable level [9, 10]. To respond to the security concern of text-dependent speaker verification systems, in some early studies, as reported in [11, 12, 13], the authors conducted the spoofing attack against an HMM-based text-prompted speaker verification system. In this paper, we make a comparative study to examine the performance of *text-dependent* and *text-independent* speaker verification systems under the same spoofing attacks, using two different voice conversion methods: *joint density Gaussian mixture model* method and *unit-selection* based method. To safeguard personal devices such as smartphones and other mobile devices, the industry has started using voice biometrics for access control of them [14]. In this paper, we would like to look into the performance of speaker verification systems under spoofing attacks on smartphones or mobile devices.

2. Database

In this work, we use the nine sessions of the first two parts of the RSR2015 database [15]. This corpus has been recorded using multiple mobile devices and smartphones as a standard benchmarking database for text-dependent speaker verification system development and evaluation. During the recording, a speaker reads 30 pass-phrases for each session of part 1 and 30 short commands for each session of part 2. The average duration of the pass-phrases is 3.2 seconds. Two non-overlapping sets of speakers are defined: a background set including 50 male

and 47 female speakers, and an evaluation set of 50 male and 47 female speakers. Speakers from the background set are reserved for training of universal background model.

During the experiments, each speaker from the evaluation set is used both as a target speaker and as an impostor against others from the same gender. Out of the 9 sessions available for each speaker, three sessions are used for enrolment (sessions 1, 4 and 7) while the six remaining sessions are used as test material. Note that enrolment and test sessions are defined so that the recording device used for test is different from the one used during the enrolment. Moreover, in order to avoid overlapping between enrolment and test sentences for the case of text-independent speaker verification, we split the 30 sentences into two groups. Pass-phrases 1 to 20 will be used as testing utterances while sentences 21 to 30 are kept for enrolment purpose. Thus, 120 utterances from each speaker are used to produce genuine and impostor trials (20 pass-phrases and 6 sessions). The statistics of the trials are presented in Table 1. Given this protocol, we note that only the genuine and impostor trials with matched pass-phrase and gender are considered.

Table 1: Statistics of the baseline and spoofing database

	Male	Female	Total
Target speakers	50	47	97
Genuine trials	5,942	5,615	11,557
Impostor trials	290,622	258,180	548,802
Converted trials	290,622	258,180	548,802

To design the spoofing attack corpus, we convert the testing segments for the impostor trials while the genuine trials are kept untouched. This allows us to focus solely on the effects of spoofing attack. We design the spoofing attack corpus by repeating the following three steps for each impostor trial:

- Establish a transformation function between the impostor and the target speaker's speech;
- Apply the transformation function to convert the testing segment of the impostor;
- Use the converted speech as testing sample for the impostor.

Thus, the number of converted trials is the same as the number of impostor trials, as shown in Table 1. Generation of the spoofing attack data, described further in Section 4, requires additional data to establish a relationship between the attacker and target speech. For this purpose, we use short commands from the part 2 of the RSR2015 database in which text material does not overlap with the 30 pass-phrases of the test data. All the 30 short commands from a single session are used as training data for voice conversion.

3. Speaker verification systems

Text-independent system has shown to be vulnerable to spoofing attack, where false alarm rate increases considerably under spoofing attacks [10, 9]. In this study, we investigate the performance of text-dependent speaker verification system under spoofing attack using text-independent speaker verification system for comparison.

The classic *Gaussian mixture model with universal background model* (GMM-UBM) [16, 17] is employed to build both text-dependent and text-independent speaker verification systems. Although more advanced techniques, for instance, joint factor analysis (JFA) [18] or i-vector PLDA [19, 20] could have been used for the spoofing attack study, as reported in our previous work [10, 9]. We consider here the well known GMM-UBM for the following reasons. Firstly, the RSR2015 text-dependent speaker recognition database consists of utterances with relatively short duration (3-second nominal duration). For short-duration training and test utterances, the conventional GMM-UBM with maximum a posteriori adaptation [21] has shown to achieve similar performance compared with JFA or PLDA. Secondly, no additional database is required for training GMM-UBM system, while JFA or PLDA systems need considerable amount of additional database to estimate the total variability, which will result in slow computation speed.

Two gender-dependent universal background models (UBM) are trained using the 50 male and 47 female speakers from the background set. Each UBM model with 64 Gaussian components is estimated using the classical expectation maximization (EM) algorithm in maximum likelihood sense.

Both text-dependent and text-independent target speaker models are obtained by adapting the UBM mean vectors with the maximum a posteriori (MAP) technique [21]. Text-dependent models differ from the text-independent ones due to the adaptation that use only constrained speech material as detailed below.

3.1. Training the text-independent speaker models

Text-independent speaker models are adapted by using three excerpts (from sessions 1, 4 and 7) of ten pass-phrases (pass-phrase 21 to 30). Thus, 30 utterances are used to adapt each target speaker model from the UBM. Note that the pass-phrases from the training material do not overlap the test utterances.

3.2. Training the text-dependent speaker models

For the case of the text-dependent speaker verification system, each pass-phrase requires its own target model. Thus three excerpts of a specific pass-phrase (from sessions 1, 4 and 7) are used for enrolment of the speaker- and pass-phrase-dependent GMM model. In this study, target speaker models are estimated for pass-phrases 1 to 20. Note that for the case of text-dependent speaker verification, each pass-phrase from a specific speaker has its own model, creating 20 target models per speaker while only one model is used in the case of text-independent speaker verification. The statistics of the models for each speaker and for all the speakers are presented in Table 2.

Table 2: Number of target models in the text-dependent and text-independent verification systems

	Text-Dependent	Text-Independent
Each speaker	20	1
All speakers	$20 \times 97 = 1{,}940$	$1 \times 97 = 97$

For both text-dependent and text-independent system, we adopt the same acoustic front-end. 12 dimensional MFCC from 27 mel-frequency filter-bands with delta and delta-delta features are extracted from original speech signal. Energy based voice activity detection is employed as post-processing to remove non-speech frames. After that, utterance level mean variance normalization (MVN) is used to make the features with zero mean and unit variance.

4. Voice conversion techniques

The task of voice conversion is to modify one speaker's (source) voice to sound like it was uttered by another speaker (target) while keeping the language content. Thus, it becomes a tool to attack both text-dependent and text-independent speaker verification systems. Voice conversion involves off-line training

and run-time conversion processes. During off-line training, a transformation function between the source and target speech is established. In the run-time conversion process, the transformation function 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.

4.1. 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 [22] and now is the mainstream approach [23, 24].

In the off-line training phase, given a parallel training data from source $\mathbf{X}$ speaker and target $\mathbf{Y}$ speaker, the source spectral vectors $\mathbf{X} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top, \ldots, \mathbf{x}_N^\top]^\top$ and target spectral vectors $\mathbf{Y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_m^\top, \ldots, \mathbf{y}_M^\top]^\top$ are aligned using dynamic time warping algorithm, where $\mathbf{x}_n \in \mathcal{R}^d$ and $\mathbf{y}_m \in \mathcal{R}^d$. The paired feature vectors are presented as $\mathbf{Z} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \ldots, \mathbf{z}_t^\top, \ldots, \mathbf{z}_T^\top]^\top$, where $\mathbf{z}_t^\top = [\mathbf{x}_n^\top, \mathbf{y}_m^\top]^\top \in \mathcal{R}^{2d}$.

Gaussian mixture model (GMM) is adopted to model the joint probability density of X and Y as:

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

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. For each Gaussian component l, $w_l^{(z)}$ is its prior probability with the constraint $\sum_{l=1}^{L} w_l^{(z)} = 1$.

The parameters of the joint density Gaussian mixture model $\lambda^{(z)} = \{w_l^{(z)}, \mu_l^{(z)}, \boldsymbol{\Sigma}_l^{(z)} | l = 1, 2, \ldots, L\}$ are estimated using the classical expectation maximization (EM) algorithm in maximum likelihood (ML) sense.

During the conversion phase, for each 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:

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

where $p_l(\mathbf{x}) = \frac{w_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l^{xx})}{\sum_{k=1}^{L} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^{xx})}$ is the posterior probability of the source vector $\mathbf{x}$ belonging to the l^{th} Gaussian.

The above transformation function is applied to the source speech feature vector sequence, after that, the converted feature vector sequence is forwarded to a speech synthesis vocoder to reconstruct audible speech signals.

4.2. Unit-selection based voice conversion

In GMM-based voice conversion, the converted speech is obtained by adopting transformation function $\hat{\mathbf{y}} = F(\mathbf{x})$ to modify the source speech. Instead of modifying the source speech, *unit-selection* based voice conversion directly makes use of the original target speech in training set to generate converted speech. The procedure of unit-selection based voice conversion is described as follows.

During the training phase, given parallel training data $\mathbf{X}$ and $\mathbf{Y}$ from source and target speakers, respectively, dynamic time warping algorithm is employed to align source $\mathbf{X}$ and target $\mathbf{Y}$ speech, in order to find the spectral vector pairs $\mathbf{Z} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \ldots, \mathbf{z}_t^\top, \ldots, \mathbf{z}_T^\top]^\top$, where $\mathbf{z}_t^\top = [\mathbf{x}_n^\top, \mathbf{y}_m^\top]^\top$ as that done in the training phase of GMM-based voice conversion. This alignment process is to guarantee the phonetic or linguistic information is not changed during the conversion phase.

In the run-time conversion phase, given a source speech feature vector $\hat{\mathbf{x}}_t$, it is paired up with the nearest source feature vector $\mathbf{x}_n$ in Euclidean distance sense. The corresponding paired target feature vector $\mathbf{y}_m$ of $\mathbf{x}_n$ is used as the converted speech vector of $\hat{\mathbf{x}}_t$. After that, the converted feature vector sequence is passed to the speech synthesis filter to reconstructed a speech signal. We note that $\mathbf{y}_m$ is the real speech frame from the original target speech, which is different from the $\hat{\mathbf{y}} = F(\mathbf{x})$ in JD-GMM conversion method.

4.3. Parameterizations for voice conversion

The speech signal, which is sampled at 16 kHz, is windowed in a 25 ms window with a 5 ms window shift. Mel-cepstral analysis [25] is employed to extract 30 dimensional mel-cepstrum coefficients (MCC) to represent the spectral envelop. During synthesis, the MCC parameters are passed to a Mel Log Spectrum Approximation (MLSA) filter [25]. In practice, the Speech Signal Processing Toolkit (SPTK) [26] is adopted to perform mel-cepstral analysis and as MLSA filter to reconstruct speech signal. Fundamental frequency values are extracted by the RAPT algorithm [27]. MCC is converted by using above conversion methods and F0 is converted by equalizing the means and variances of source and target speakers in log-scale.

5. Experimental results and discussion

To evaluate the vulnerability of the speaker verification, we employ the *equal error rate* (EER) and *MinDCF* (adopting the cost parameters in the NIST SRE 2006 plan) measures.

We employ the JD-GMM and unit-selection methods to generate spoofing data, as discussed in Section 2. We generate the same number of original impostor trials, JD-GMM converted trials, and unit-selection converted trials. There are 290,622 and 258,180 trials for male and female speakers, respectively. When calculating EER, the baseline test involves a mix of genuine trials and original impostor trials, while the converted voice test involves a mix of genuine trials and converted trials. The actual numbers of trials are shown in Table 1. In this way, we have the same number of trials for baseline test, JD-GMM converted voice test, and unit-selection converted voice test.

The EER and MinDCF results of the text-dependent and text-independent speaker verification systems under spoofing attack are presented in Table 3. For the text-independent speaker verification system, the EER increases signicantly under spoofing attacks from 17.17% to 30.20% and 27.53% for female speakers by unit-selection and JD-GMM conversions, respectively, and the MinDCF also increases. This confirms our previous finding using several text-independent verification systems including the classic GMM-UBM system and the state-of-the-art JFA or PLDA system [10, 9]. However, for the text-dependent speaker verification, the EER of female trials drops from 4.79% to 2.39% and 1.84% for unit-selection and JD-GMM based conversions, respectively, and the MinDCF also drops. We note that the target speaker model in the text-dependent speaker verification system is estimated using the matched pass-phrase utterances. Thus, the phonetic or linguistic information is already taken into account. As the two voice conversion conversion methods only focus on converting the

Table 3: Performance of *text-dependent* and *text-independent* speaker verification systems under *spoofing*.

	Equal error rates (EER %)				100 × MinDCF			
Voice conversion	Text-Dependent		Text-Independent		Text-Dependent		Text-Independent	
	Male	Female	Male	Female	Male	Female	Male	Female
None (Baseline)	6.62	4.79	15.32	17.17	3.25	2.81	6.85	7.71
Unit-selection	4.85	2.39	27.30	30.20	2.76	1.55	10.00	9.92
JD-GMM	4.51	1.84	25.87	27.53	2.55	1.19	9.99	9.88

spectral envelop, the duration and prosody information, which are important factors in phonetic or linguistic information, are kept the same as the impostor's. Thus, the phonetic or linguistic knowledge enables the text-dependent system to discriminate converted speech from the original target speech.

We observe that the text independent speaker verification system gives a higher EER under spoofing attacks, which suggests that the spoofing attacks increase the log likelihood scores of imposter trials. The increase of log-likelihood scores has compromised the decision of the verification system. On the other hand, text-dependent speaker verification system gives lower EER under conversion spoofing attack. The phonetic or linguistic discriminative information lowers the log-likelihood scores of converted impostor trials, and strengthens the ability of making correct decision. To verify our assumption, we present the log-likelihood scores distribution of impostor trials before and after conversion in Fig. 1 and 2 from text-dependent and text-independent system, respectively. From both score distributions, it shows that mean score under unit-selection conversion spoofing attack is higher than that under JD-GMM conversion attack, which is consistent with the EER and MinDCF results.

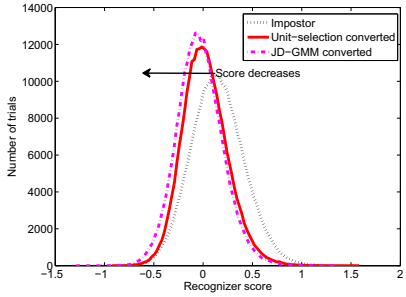

Figure 1: Impostor score distribution of text-dependent verification system before and after conversion spoofing attacks

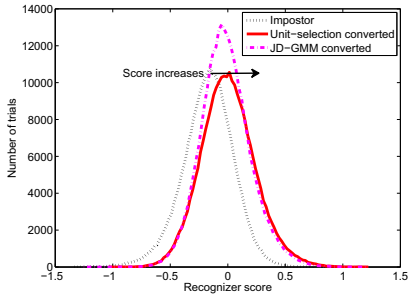

Figure 2: Impostor score distribution of text-independent verification system before and after conversion spoofing attacks

Comparing with the two conversion methods, unit-selection based conversion always gives higher EER and MinDCF than JD-GMM based conversion. We note that unit selection based conversion directly select target frames to compose the converted speech. It is easy to understand that it generates speech more similar to target speaker than JD-GMM based conversion does, the latter employs a transformation function to map the source speech to the target space.

From the view of voice conversion, unit-selection based conversion directly selects target frames to make the converted speech. Thus, it is able to generate speech more similar as target speaker than JD-GMM based conversion, which employs transformation to shift the source speech to the target space.

In real application, the decision threshold of a speaker verification system is fixed and the threshold is based on the baseline database without spoofing. The speaker verification system is unaware of spoofing attacks. Thus, the spoofing attacks may affect the *false acceptance rate* (FAR) considerably. To assess the spoofing attack effect, we set the decision threshold to the EER point on the baseline corpus without conversion, and then examine the FAR under spoofing attacks. The FARs under spoofing attacks simulated by unit-selection and JD-GMM conversions are presented in Table 4. The FAR of text-independent system for female speakers is increased from 17.17% to 44.40% and 39.28% by unit-selection and JD-GMM conversions, respectively. However, for text-dependent system, FAR is decreased from 4.79% to 1.19% and 0.73% of females by unit-selection and JD-GMM conversions, respectively.

Table 4: Spoofing attack effect on false acceptance rates (FAR, %). The verification decision threshold is set to the EER point on the baseline corpus.

	Text-Dependent		Text-Independent	
Voice conversion	Male	Female	Male	Female
No conversion	6.62	4.79	15.32	17.17
Unit-selection	3.44	1.19	42.56	44.40
JD-GMM	2.88	0.73	39.22	39.28

From the EER, MinDCF and FAR results, we observe that text-dependent speaker verification system works much better than text-independent speaker verification system under spoofing attacks. We note that there are no original impostor trials in the spoofing database. The FAR results show that the attacker still has chances to break the text-independent system.

6. Conclusions

In this study, we examine the vulnerability of speaker verification by comparing the performance of text-dependent and text-independent speaker verification systems under spoofing attacks, which are simulated by JD-GMM and unit-selection based voice conversion methods. The experimental results show that the phonetic or linguistic information is helpful in discriminating converted speech from natural speech and makes the text-dependent verification system more robust against spoofing attack than text-independent system.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

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

[3] 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.

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

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

[6] 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*.

[7] 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.

[8] Q. Jin, A. Toth, A. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?" in *ICASSP 2008*.

[9] 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*.

[10] 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 *ICASSP 2012*.

[11] 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*.

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

[13] 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*.

[14] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter, Feb. 2013*.

[15] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases," in *INTERSPEECH 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, 2000.

[17] H. Li and B. Ma, "Techware: Speaker and spoken language recognition resources," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 139–142, 2010.

[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

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

[20] 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.

[21] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241–1269, 2000.

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

[23] 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.

[24] 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.

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

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

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