Voice conversion and spoofing attack on speaker verification systems

Zhizheng Wu*, Haizhou Li*†
*Nanyang Technological University, Singapore
†Institute for Infocomm Research, Singapore
wuzz@ntu.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—We design speaker verification system to automatically accept or reject the claimed identity of a speaker. Recently, we have made major progress in speaker verification which leads to mass market adoption, such as in smartphone and in online commerce for user authentication. The major concern when deploying speaker verification technology is whether a system is robust against spoofing attacks. Speaker verification studies provided us a better insight into speaker characterization, which has contributed to the progress of voice conversion technology. Unfortunately, voice conversion has become one of the most easily accessible techniques to carry out spoofing attack, therefore, presents a threat to speaker verification systems. In this paper, we will briefly introduce the spoofing attack studies under different conditions with a focus on voice conversion spoofing attack. We will also discuss anti-spoofing attack measures for speaker verification.**

I. Introduction

A large number of physical or behavioural attributes, which are distinctive, measurable characteristics to describe human individuals, have been investigated for biometric recognition. Speaker verification is among the most popular biometrics in smartphone [1] or telephony applications where voice service is provided. It is also called voice biometrics. The task of speaker verification is to automatically accept or reject a claimed identity based on a speech sample. Fig. 1 is an illustration of a typical speaker verification system.

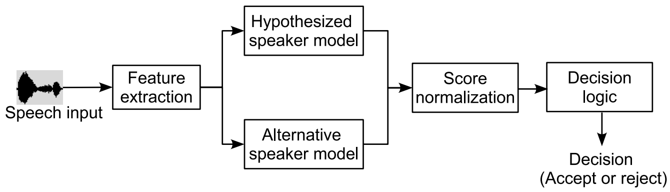

Fig. 1. Diagram of a speaker verification system.

Just like any other means of biometrics, a speaker verification system is expected to be not only accurate for regular users, but also secure against spoofing attacks. As discussed in [2], possible spoofing attack happens at two points: sensor level and transmission of the sensed signal. At the sensor level, an adversary, that we call an impostor, could deceive the system by impersonating someone at the microphone, and at the transmission time when the acquired voice signal could be replaced by a synthetically generated signal or imitated voice. In general, spoofing attack is to use a falsifying speech signal as system input (See Fig. 1) for feature extraction and verification, therefore, present a threat to speaker verification systems.

As digital recording has become widely accessible, *replay attack* is the simplest method to spoof a speaker verification system. Replay attack involves repetition of a pre-recorded speech sample or a sample created by concatenating basis speech segments from a given target speaker. Indeed, replay attack has been shown to be an effective way to spoof text-independent recognizers which do not utilize any linguistic constraints [3], [4]. However, the replay technique is not flexible in generating specific utterances as required by text-dependent speaker verification systems.

Aside from replay attack, *human voice mimicking* or *impersonation* has also received considerable attention [5], [6], [7]. As impersonation requires special skills, it is difficult to judge its effectiveness as a spoofing technique. Partial evidence, however, suggests that humans are most effective in mimicking speakers with "similar" voice characteristics to their own, while impersonating an arbitrary speaker appears challenging [5]. Professional voice mimics, often voice actors, tend to mimic prosody, accent, pronunciation, lexicon and other high-level speaker traits, rather spectral cues used by automatic systems. Therefore, human voice mimicking is not considered as a cost-effective adversary to speaker verification systems.

Speech synthesis represents a much more genuine threat. Due to the rapid development of *unit selection* [8], *statistical parametric* [9] and *hybrid* [10] methods, speech synthesis systems are now able to generate speech with a certain speaker's voice characteristics, such as spectral cues, and acceptable quality. In early studies [11], [12], [13], vulnerability of text-prompted *hidden Markov model* (HMM) based speaker verification was examined using a small database of 10 speakers. More recently, [14] used a flexible adapted HMM-based speech synthesis system to simulate spoofing attacks against text-independent recognizer on a corpus of around 300 speakers. Even though HMM-based synthesis poses a threat, a lot of training speech (usually one hour or more) is needed to train the speech synthesis system. Even for adapted HMM-based speech synthesis system, one needs additional speakers' data to train an average voice model for target speaker adaptation [15]. Therefore, it is expensive for attackers to conduct spoofing attack using a HMM-based speech synthesis system.

Different from replay attack, human voice mimic and text-to-speech, *voice conversion* is to modify one speaker's (source) voice to sound like it was pronounced by another speaker (target) without changing the language content. While keeping the language content unchanged, the conversion technique works in two ways, one is to change the source voice to sound differently - to disguise oneself; the other is to change the source voice to a target voice - to mimic someone else. As real-time voice conversion not only is possible, but also offers voice quality and characteristics that even human ears are hard to distinguish, it presents a genuine threat to both text-dependent and text-independent speaker verification systems.

In summary, human voice can be seen to have three attributes, the language content, the spectral pattern, and the prosody. The individuality of human voice is described by the spectral patterns, called voice quality or timbre, and by the prosodic patterns carried by the speech. Professional voice mimicking typically modifies the prosodic patterns while voice conversion modifies the spectral patterns. As it is more reliable to characterize speakers by their spectral cues [16], most of the state-of-the-art speaker verification systems are built to detect the difference of spectral patterns. In this paper, we will focus on the conversion spoofing attack, and review the most recent research works on voice conversion, speaker verification, spoofing attack and anti-spoofing attack techniques.

The rest of this paper is organized as follows. In Section II, we will briefly review the state-of-the-art speaker verification techniques, and in Section III, an overview of voice conversion techniques is presented. Spoofing attack and anti-spoofing attach studies are reviewed in Section IV and Section V. The paper is concluded in Section VI.

II. Speaker verification techniques

The objective of a speaker verification system is to automatically accept or reject a claimed identity S of one speaker based on just the speech sample $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t, \cdots, \mathbf{o}_T\}$ [16]. This verification process is illustrated in Fig. 1 and is formulated into statistical form as:

$$\Lambda(\mathbf{O}) = \frac{p(\mathbf{O}|\lambda_H)}{p(\mathbf{O}|\lambda_{\widetilde{H}})}, \tag{1}$$

where λ_H is the model parameters of hypothesis H that the speech sample $\mathbf{O}$ is from speaker S, and $\widetilde{H}$ is a alternative hypothesis that the speech sample is not from the claimed identity S. The likelihood ratio (or likelihood score) $\Lambda(\mathbf{O})$ is used to decide which hypothesis, H or $\widetilde{H}$, is true based a pre-defined threshold.

In practice, there are two kinds of speaker verification systems: *text-independent* speaker verification (TI-SV) and *text-dependent* speaker verification (TD-SV) systems. TD-SV assumes cooperative speakers and requires the speaker to speak fixed or randomly prompted utterances, while TI-SV allows the speaker to speak freely during both enrolment and verification. In general, both TI-SV and TD-SV systems adopt the same feature extraction techniques, that we call the front-end. In this section, we will briefly describe the stat-of-of-art speaker verification systems.

A. Feature extraction

Studies show that there are three level of features characterize the individuality of speakers: high level, spectro-temporal and short-term features [16]. As speech signal is not stationary, shifting windows are usually applied to divide the speech signal into short-term overlapping segments with around 20 to 30 msec before extracting features.

High level features, which involves phoneme, accent, pronunciation, etc, are robust against noise, but different to be extracted. Usually, automatic speech recognition is required to extract high level features. Spectro-temporal features involve prosodic, temporal modulation features, etc. Short-term features are extracted from fixed size speech frames. Among the three level of features, studies have shown that short-term features are the most cost-effective in practice.

Mel-frequency cepstral coefficient (MFCC), linear predictive cepstral coefficient (LPCC), and perceptual linear prediction (PLP) are the most popular short-term features. Usually we also include delta and delta-delta coefficients of these short-term features to take speech dynamics into consideration.

B. Speaker modeling

Text-independent speaker verification systems focus on modeling the feature distribution of target speaker. Gaussian mixture model (GMM) has been used intensively to model the feature distribution. GMM-UBM is the classical method in early speaker verification systems [17]. Maximum likelihood and maximum a posteriori training have been adopted in such system. Tens or hundreds hours speech gathered from a large number of speakers are first employed to estimate a speaker independent universal background model (UBM), and then the target speaker model is obtained by adapting the universal background model with several minutes speech from the target speaker. The target speaker GMM model and the UBM model are used as hypothesised speaker model and alternative speaker model, respectively.

Speaker verification is a two class classification problem, discrimination between target speech model and background model is important. To allow discriminative training of generative model, for example GMM-UBM model, support vector machine (SVM) is combined with GMM-UBM framework, where GMM mean supervectors are used as feature in SVM classifier [18]. In the context of SVM modeling, Nuisance Attribute Projection (NAP) [19], [20] and within-class covariance normalization (WCCN) [21] techniques are proposed for channel compensation.

In recent research, a generative model, joint factor analysis (JFA), is proposed for intersession and speaker variability compensation [22], [23]. JFA is a latent variable model to explicitly model channel and speaker variability jointly, which use a large number of additional data to estimate both speaker and channel variabilities. The estimated speaker and channel variabilities are employed to estimate the speaker identity vector and channel vector during enrolment and verification.

Similar to JFA method, probabilistic linear discriminant analysis (PLDA), which a generative model, employs i-vector rather than GMM supervectors as the basis for estimating factor loadings [24]. PLDA models speaker and channel variabilities within i-vector space. An i-vector is a low-dimensional set of factors to represent speaker and channel information via factor loadings (total variability) [25]. Similar as JFA method, a large number of additional data is used to estimate the total variability matrix (factor loadings), which represent both speaker and channel variabilities. These additional data is also adopted to estimate the speaker and channel variabilities in i-vector space.

In practice, it is not enough to build just one single strong recognizer. It is a general practice to fuse multiple sub-systems into one as a mixture-of-expert. This is based on the assumption that individual classifiers are able to capture different aspects of the speech signal, thus providing complimentary information for each others. Each individual classifier can involve different kinds of features or different level of features, and can also employ different modeling techniques. While fusion usually takes place at score level across subsystems [26], [27], [28], there are also ways to fuse the features or speaker models [28].

Different from text-independent speaker verification systems, text-dependent systems not only model the feature distribution, but also take the linguistic information and temporal into consideration. Therefore, hidden Markov model (HMM) is one of the base models. All the techniques developed for text-independent systems can be easily transferred to the text-dependent ones by using HMM to learn the temporal information.

III. Voice conversion techniques

Voice conversion is closely related to speaker verification. The former analyze and synthesize the voice characteristics of speakers, while the latter distinguishes one from another. Mathematically, voice conversion is a process to learn a conversion function $\mathcal{F}(\cdot)$ from source $\mathbf{X}$ and target speech $\mathbf{Y}$, and to apply this conversion function to a testing source speech signal $\mathbf{X}$ in order to generate a converted speech signal $\widetilde{\mathbf{Y}}$. This process is formulated as follows:

$$\widetilde{\mathbf{Y}} = \mathcal{F}(\mathbf{X}). \tag{2}$$

Similar to speaker verification, voice conversion also operates on features which character speaker individuals, such as formant [29], spectrogram [30], [31], [32], fundamental frequency [33], [34], [35], [36], duration [34], [37] and so on. As spectrogram contains more speaker identity information, most of the research works focus on spectral conversion, in this section, we will give a overview of the spectral conversion.

Generally, spectral conversion involves training and conversion phases as illustrated in Fig. 2. During training phase, features, which characterize speaker's individuality, are first extracted from both source and target speech signal. Then, each source feature vector is paired up with one target feature vector, which is called frame alignment, to guarantee the language content is kept the same before and after conversion. The frame alignment is usually done by dynamic time warping for parallel data [38], or through some advanced frame alignment techniques for non-parallel data [39]. Finally, a conversion function can be estimated from the source-target frame pairs. In the conversion phase, the conversion function estimated in the training phase is employed to the features extracted from source speech, and then the converted feature vector sequence is passed to a synthesis filter to reconstruct audible speech signal. It is apparent that feature extraction and estimation of the conversion function are the two most important processes in a voice conversion system.

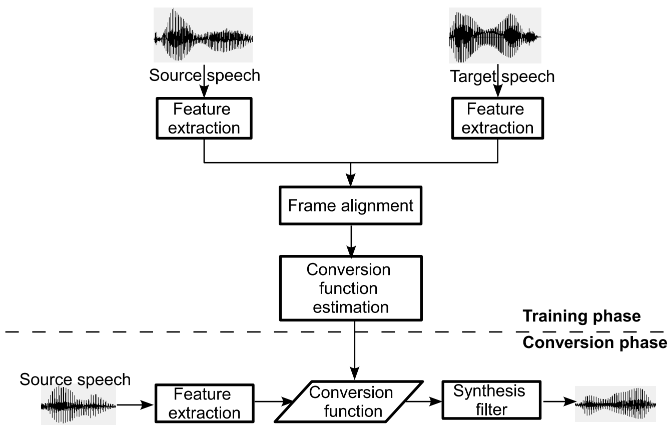

Fig. 2. Diagram of the voice conversion system.

A. Feature extraction

To extract feature for representing speech signal, a speech production model, such as source-filter model and sinusoidal model, is employed to separates speech signal into mutually independent representation components. The same speech production model is also employed to reconstruct speech signal. The following three models/systems are widely used in voice conversion task for speech analysis and synthesis:

a) STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) system [40] is based on source-filter model
b) Harmonic plus noise model (HNM) [41] separates a speech signal into a harmonic part and a noise part based on a pitch synchronous decomposition.
c) Linear prediction (LP) model assumes current speech sample is predicted as a linear combination of its past p samples, where p is the order of LP coefficients.

Above speech production models are able to extract spectrum effective and synthesize high quality speech signal. To handle the relatively high dimensional spectrum, parametric representations or features are extracted to represent the spectrum. The most popular two features used in voice conversion are listed as follows:

a) Mel-cepstral coefficient (MCC) [32], [42], [43]: MCC is obtained by applying mel-cepstral analysis technique [44] on the magnitude spectrogram and keeping 24 coefficients as the feature.

b) Line spectral frequency (LSF) [45], [46]: LSF features have good quantization and interpolation properties, and have been successfully applied to speech coding [47]. LSFs are closely related to formants, which represent speaker identity.

The synthesis step of voice conversion is similar to that of general speech synthesis, therefore, features or spectral representations used in speech synthesis can be adopted in a voice conversion task.

B. Conversion function

In the past decades, a large number of spectral conversion methods have been proposed. These methods are roughly grouped into three categories: generative methods, transmutative methods, and exemplar-based methods. In the generative methods, the converted speech signal is generated from some parametric model, and in the transmutative methods, small number of parameters are employed to control the shape of the spectrum or spectral envelop, while in the exemplar-based methods, the converted speech is constructed by using the segments of original speech from the target speaker.

1) Generative methods: The first generative method, vector quantization (VQ), is simple and straightforward [48]. With VQ, a codebook of paired source-target frame vector is built during the training phase to present the relationship between source and target speech. Although VQ method is able to capture speaker identity information for same language content, the frame-to-frame discontinuity problem caused by the codebook conversion function affects the converted speech quality considerably.

To overcome the frame-to-frame discontinuity arising from VQ method, segmental codebook [49], and fuzzy vector quantization [50] are proposed. Gaussian mixture model (GMM) based methods including joint density GMM [30] and source GMM [31] are also studied to implement local linear transformation functions based one Gaussian mixture model. These weighted local linear transformation functions are able to generate smooth spectral trajectories. In addition, maximum likelihood spectral trajectories generation algorithm is proposed in [32] based on joint density Gaussian mixture model (JD-GMM) to include local temporal information for smooth spectral trajectories. Similarly, trajectory hidden Markov model is proposed in [43] to include local dynamic information explicitly.

Although JD-GMM method becomes one of the most popular methods benefiting from a well grounded probabilistic formulation, over-smoothing [46], [51], [52] and over-fitting [53], [54] have been reported. To address these problems, a number of methods have been proposed. In [53], partial least square regression method has been proposed to avoid over-fitting problem when the parallel training data is limited, local linear transformation method is implemented by using nearby training data not all the training data to estimate the transformation function, mixture of factor analyzers [55] and noisy channel model [56] methods have been proposed to utilize additional data to improve the performance of the conversion function.

Above methods assume that the source and target features have linear relationship. Another idea is to assume that the source and target speech features have non-linear relationship, that leads to another group of methods, such as artificial neural network [29], [57], support vector regression [58], kernel partial least square [59], and conditional restricted Boltzmann machine [60].

2) Transmutative methods: In the statistical parametric methods, the conversion function is formulated from the parametric representations of the spectrum without any physical principles, and the statistical averaging effect, which reflects the central tendency of speech features, will introduce over-smoothing [51], [32], [61] Different from data-driven generative methods, which operate on the low-dimensional parametric representations of spectrum, frequency warping methods aim to warp the frequency axis of the amplitude spectrum [62], [63], [64], [65], [66]. Therefore, frequency warping or vocal tract length normalization (VTLN) methods can keep more spectrum details and produce high quality speech signal. Although frequency warping methods are able to produce high quality converted speech, the similarity between converted and target speech of frequency warping methods is not as good as generative methods as reported in [65].

3) Exemplar-based methods: In general, generative methods and transmutative methods are to modify the speaker characteristics. Different from these methods, exemplar-based methods utilize original target speaker's feature vectors to construct the converted speech. In [67], [68], [69], unit-selection method is implemented, where a single frame is used as the base unit. To include temporal information for avoiding frame-to-frame discontinuity, in [69], multiple-frame speech segment (exemplar) is employed as base unit and a temporal window is adopted to deal with overlapping frames for temporal continuity. In the unit-selection methods, usually more parallel training data is required to cover more unit patterns.

Aside from the unit-selection implementation, another exemplar-based method is based on non-negative matrix factorization. In [61], non-negative spectrogram factorization and non-negative spectrogram deconvolution are employed to use relative high dimensional spectrum directly for synthesizing speech signal. Each converted spectrogram is represented as a linear combination of source spectrums or spectrogram segments.

IV. Spoofing attack studies

As shown in Eq.(1) and Eq.(2), voice conversion modifies the input $\mathbf{O}$ to speaker verification systems, that presents a threat to speaker verification systems. Fig. 3 illustrates a general framework for voice conversion spoofing attack study.

As spoofing attack study involves both voice conversion and speaker verification, we look into three areas:

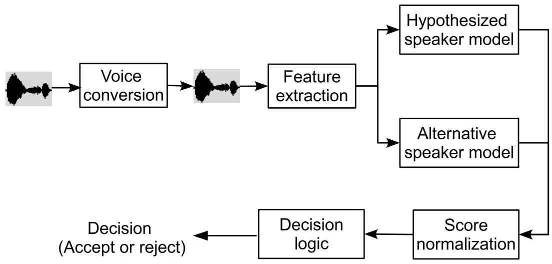

Fig. 3. Illustration of spoofing attack study of speaker verification.

a) The practicality and effectiveness to use voice conversion to make a spoofing attack.
b) The vulnerability of speaker verification recognizers under voice conversion attack.
c) The design of a realistic data set for voice conversion attack experiments.

In speaker verification, the decision of a test sample or a trial, belongs to one of the four categories (See Table I). If the test sample and the hypothesized model are from the sample speaker, we call it a genuine test, otherwise, an imposter test. Equal error rate (EER) is a common evaluation measure to balance false alarm and miss detection. The optimization of system is to reduce EER, however, spoofing attack is trying to increase false alarm errors. To measure the effectiveness of voice conversion spoofing and evaluate vulnerability of speaker verification, false acceptance rate (FAR) and EER good performance indicators. In a typical voice conversion spoofing scenario, an impostor attempts to deceive a system by using converted voice to represent a target speaker. We simulate such scenario by synthesizing target speaker's voice using voice conversion. We expect the impostor attempts to increase the false alarm, therefore, higher false acceptance rate and equal error rate of the systems.

TABLE I
FOUR CATEGORIES OF TRIAL DECISIONS IN SPEAKER VERIFICATION.

	Decision	
	Accept	Reject
Genuine	Correct acceptance	Miss detection
Impostor	False alarm	Correct rejection

A. Database design

To provide an objective assessment of system performance under voice conversion attack, we need to set up an evaluation database that allows us to compare the system performance with that without attacks. Here we use National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE) 2006 core task, 1conv4w-1conv4w, as a case study. More details can be found in [70].

The statistics of trials and speakers used in the voice conversion attack study in [70] are presented in Table II. To design the spoofing database, we first select impostors and corresponding target genuine speakers. Then, we use the *3conv4w* and *8conv4w* training sections in the NIST SRE 2006 database to estimate the conversion function for each impostor-target speaker pair. Finally, we process the each impostor's testing utterances in the trial list using the pre-estimated conversion function. Hence, the number of converted trials is the same as that of the original impostor trials, and the genuine trials are kept unchanged as in original test. Therefore, we are able to compare the original results with spoofing results by using the same number of trials. This ideal setup may be different from practice, where impostor trials and converted trials are mixed together.

The database design for both text-dependent (-constraint) and text-independent speaker verification scenarios is presented in [71]. We note that in [71], there are strong constraint on both training of speaker verification recognizers and training of conversion function.

TABLE II
SUBSET OF NIST SRE 2006 CORE TASK IN THE SPOOFING ATTACK EXPERIMENTS [70] (VC=VOICE CONVERSION).

	Standard speaker verification	Spoofing attack
Unique speakers	504	504
Genuine trials	3,978	3,978
Impostor trials	2,782	0
Imposter trials (via VC)	0	2,782

B. Experiments

To evaluate the vulnerability of speaker verification systems under voice conversion attack. A number of studies have been conducted. Table III summarizes the voice conversion spoofing attack studies. In [78], voice conversion is done by mapping the impostor's vocal tract information towards that of the genuine speaker, using frequency warping technique. The experiments conducted on NIST SRE 2005 database show that the equal error rate (EER) is increased from around 10% to over 60% when all the impostor samples are converted towards the genuine speaker. Using same voice conversion method, in [72], the authors evaluate the GMM-UBM verification system on both NIST SRE 2005 and NIST SRE 2006 databases. The EERs are increased from 8.54% and 6.61% to 35.41% and 28.07% on NIST SRE 2005 and 2006 databases, respectively. With same database and same conversion method, in [73], the vulnerability of GMM-UBM and JFA systems are compared under conversion spoofing attack. The experimental results show that the EERs are increased from 8.5% and 4.8% to 32.6% and 24.8% of GMM-UBM and JFA systems, respectively. In [74] and [70], multiple state-of-the-art speaker verification systems are compared using the same database and JD-GMM voice conversion method to simulate spoofing attack. The experimental results show that the spoofing attacks increase the EER more than two times over that of the baseline for all the text-independent systems.

Different from above studies, which only focus on text-independent (TI) systems. In [75], [71], a comparison of text-dependent (TD) and text-independent systems under voice conversion spoofing is conducted. In [75], two TI recognizers,

TABLE III
SUMMARY OF VOICE CONVERSION SPOOFING ATTACK STUDIES (TI=TEXT-INDEPENDENT RECOGNIZER; TC=TEXT-CONSTRAINT; TD=TEXT-DEPENDENT).

				Baseline	Spoofing	
Conversion method	Database	TI or TC or TD	Recognizer	EER (%)	EER (%)	FAR (%)
Frequency warping [72]	NIST SRE 2005	TI	GMM-UBM	8.54	35.41	N. A.
Frequency warping [72]	NIST SRE 2006	TI	GMM-UBM	6.61	28.07	N. A.
Frequency warping [73]	NIST SRE 2005	TI	GMM-UBM	8.50	32.60	N. A.
Frequency warping [73]	NIST SRE 2005	TI	JFA	4.80	24.80	N. A.
JD-GMM [74]	NIST SRE 2006	TI	GMM-UBM	7.63	24.99	N. A.
JD-GMM [74]	NIST SRE 2006	TI	VQ-UBM	7.56	22.62	N. A.
JD-GMM [74]	NIST SRE 2006	TI	GMM-SVM	3.74	12.58	41.54
JD-GMM [74]	NIST SRE 2006	TI	JFA	3.24	7.61	17.33
Unit-selection [70]	NIST SRE 2006	TI	JFA	3.24	11.58	32.54
JD-GMM [70]	NIST SRE 2006	TI	PLDA	2.99	6.77	19.29
Unit-selection [70]	NIST SRE 2006	TI	PLDA	2.99	11.18	41.25
Frequency warping [75]	WF corpus [76]	TI	I-vector	1.60	8.80	29.00
Frequency warping [75]	WF corpus [76]	TI	GMM-NAP	1.10	3.40	38.00
Frequency warping [75]	WF corpus [76]	TD	HMM-NAP	1.00	2.90	36.00
JD-GMM [71]	RSR2015 [77]	TI	GMM-UBM	15.32	25.87	39.22
Unit-selection [71]	RSR2015 [77]	TI	GMM-UBM	15.32	27.30	42.56
JD-GMM [71]	RSR2015 [77]	TC	GMM-UBM	6.62	4.51	2.88
Unit-selection [71]	RSR2015 [77]	TC	GMM-UBM	6.62	4.85	3.44

I-vector and GMM-NAP, and one TD recognizer, HMM-NAP are adopted. The experimental results show voice conversion increases EER for both TI and TD recognizers. Differently, in [71], a text-constraint database is employed for text-dependent study, and only matched transcript trials are used in this study. Hence, we can treat text-constraint (TC) GMM-UBM system as a TD system. In addition, a constraint part of the database, which has only digits, is used to train the voice conversion function. The experimental results show voice conversion increases EER for TI system, but surprisingly reduces EER for TD system.

In general, above experimental results suggest that advance speaker verification systems, such as JFA and PLDA, are more vulnerable than lightweight GMM-UBM and VQ-UBM techniques under voice conversion spoofing attack. Across the board, the performance of all systems is compromised to an unacceptable level under the attacks. On the other hand, unit-selection conversion method is more effective than JD-GMM method in deceiving speaker verification systems, as unit-selection method directly uses target speaker's voice segment to synthesize the spoofing voice.

V. ANTI-SPOOFING ATTACK STUDIES

As shown in Section IV, the performance in terms of equal error rate of speaker verification systems, even the state-of-the-art speaker verification systems, is degraded considerably under voice conversion spoofing attack. The research is addressing the problem in two ways:

a) Improve the performance of the speaker verification recognizers. As shown in [74], [73], [70], advanced algorithms, such JFA and PLDA, are more robust against spoofing attack.
b) Design converted speech detectors that is able to tell synthesized voice from naturally produced voice.

We have seen successful techniques that detect artifacts introduced during the voice conversion or synthesis process. Fig. 4 is an example of incorporating a converted speech detector as an explicit countermeasure against spoofing attack.

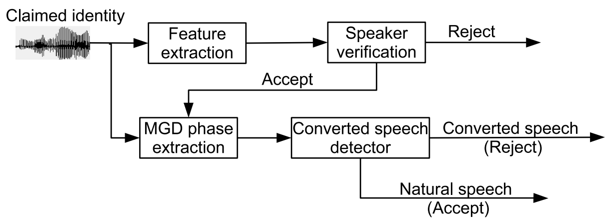

Fig. 4. Diagram of speaker verification with an anti-spoofing synthetic speech detector [70] (MGD = modified group delay).

To design the converted speech detector, in [79], [70], cosine normalized phase (cos-phase) and modified group delay phase (MGD-phase) features are proposed based on the fact that the original phase information is lost during the analysis-synthesis step in some vocoders. The experiments using NIST SRE 2006 database show that the detection EER of 5.95% and 2.35% can be obtained by using cos-phase and MGD-phase, respectively.

In [80], high level features, which is extracted over a long speech context, is employed to capture the change of speech dynamics and to distinguish converted speech from natural human speech.

As current analysis-synthesis techniques for extracting features and synthesizing speech signal operate on short-term frame level (5 msec to 15 msec), some artifacts are introduced in the temporal domain. Therefore, temporal magnitude or phase modulation features are able to detect converted speech which utilizes vocoding techniques [81]. Comparing with MGD-phase, temporal modulation feature reduces EER from 1.25% to 0.89% on Wall Street Journal (WSJ0+WSJ1) database. Similarly, pair-wise distance between consecutive feature vectors is adopted to present short-term speech variability and detect converted speech.

In [70], the converted speech detector, which employs MGD-phase, is combined with two speaker verification systems. The false acceptance rate of GMM-JFA system is reduced from 17.36% and 32.54% to 0.0% and 1.64% for GMM and unit-selection conversion spoofing, respectively, and FAR of PLDA system is reduce from 19.29% and 41.25% to 0.0% and 1.71% for GMM and unit-selection conversion spoofing, respectively.

VI. Conclusion

In this paper, we present an overview of spoofing, anti-spoofing attack and related techniques. Due to rapid development of speaker verification techniques, speaker verification systems have been deployed into real applications, such as smartphone [1]. At the same time, voice conversion techniques also progress quickly. Therefore, the countermeasures for spoofing attacks become an important part of system deployment. The current studies on anti-spoofing attack are very preliminary because the results are reported only on selected techniques. Comprehensive studies on the effects of different synthesis and vocoding techniques and different speaker recognition regimes are expected in the near future. In INTERSPEECH 2013, a special session on 'Spoofing and countermeasures for automatic speaker verification' is organized for the first time, which shows the increasing importance and attention of this research topic given by the academia and industry.

References

[1] Kong Aik Lee, Bin Ma, and Haizhou Li, "Speaker verification makes its debut in smartphone," in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter,*, February 2013.

[2] Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin, "Speaker verification security improvement by means of speech watermarking," *Speech communication*, vol. 48, no. 12, pp. 1608–1619, 2006.

[3] Johan Lindberg, Mats Blomberg, et al., "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proc. the European Conference on Speech Communication and Technology*, 1999.

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

[5] Yee Wah Lau, Michael Wagner, and Dat Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[6] Mireia Farrús, Michael Wagner, Jan Anguita, and Javier Hernando, "How vulnerable are prosodic features to professional imitators?," in *The Speaker and Language Recognition Workshop (Odyssey 2008)*, 2008.

[7] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proc. INTERSPEECH*.

[8] Andrew J Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996.

[9] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] Yao Qian, Frank K Soong, and Zhi-Jie Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1-2, pp. 280–290, 2013.

[11] Takashi Masuko, Takafumi Hitotsumatsu, Keiichi Tokuda, and Takao Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999.

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

[13] Takayuki Satoh, Takashi Masuko, Takao Kobayashi, and Keiichi Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eurospeech*, 2001.

[14] Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[15] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

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

[17] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

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

[19] Alex Solomonoff, William M Campbell, and Ian Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*.

[20] Lukas Burget, Pavel Matejka, Petr Schwarz, Ondrej Glembek, and Jan Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[21] Andrew O Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. ICSLP*, 2006.

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

[23] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

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

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

[26] V. Hautamaki, T. Kinnunen, F. Sedlak, Kong Aik Lee, Bin Ma, and Haizhou Li, "Sparse classifier fusion for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1622–1631, 2013.

[27] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boril, and John HL Hansen, "CRSS systems for 2012 nist speaker recognition evaluation," in *Proc. ICASSP*, 2013.

[28] Mitchell McLaren, Nicolas Scheffer, Martin Graciarena, Luciana Ferrer, and Yun Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. ICASSP*, 2013.

[29] M. Narendranath, H.A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[30] Alexander Kain and Michael W Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998.

[31] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

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

[33] B Gillet and S King, "Transforming F0 contours," in *Proc. Eurospeech*, 2003.

[34] Chung-Hsien Wu, Chi-Chun Hsia, Te-Hsien Liu, and Jhing-Fa Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, 2006.

[35] Elina E Helander and Jani Nurminen, "A novel method for prosody prediction in voice conversion," in *ICASSP*, 2007.

[36] Zhi-Zheng Wu, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. INTERSPEECH*, 2010.

[37] Damien Lolive, Nelly Barbot, and Olivier Boeffard, "Pitch and duration transformation with non-parallel data," *Proc. Speech Prosody*, pp. 111–114, 2008.

[38] Elina Helander, Jan Schwarz, Jani Nurminen, Hanna Silen, and Moncef Gabbouj, "On the impact of alignment on voice conversion performance," in *Proc. INTERSPEECH*, 2008.

[39] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[40] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[41] Yannis Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[42] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *ICASSP*, 2007.

[43] Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

[44] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation.," in *Proc. ICSLP*, 1994.

[45] Elina Helander, Jani Nurminen, and Moncef Gabbouj, "LSF mapping for voice conversion with very small training sets," in *Proc. ICASSP*, 2008.

[46] Victor Popa, Hanna Silen, Jani Nurminen, and Moncef Gabbouj, "Local linear transformation for voice conversion," in *Proc. ICASSP*, 2012.

[47] Frank Soong and Biing-Hwang Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984.

[48] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988.

[49] Levent M Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[50] K Shikano, S Nakamura, and M Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proc. IEEE International Sympoisum on Circuits and Systems*, 1991.

[51] Yining Chen, Min Chu, Eric Chang, Jia Liu, and Runsheng Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proc. Eurospeech*, 2003.

[52] Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Yih-Ru Wang, and Sin-Horng Chen, "A study of mutual information for GMM-based spectral conversion," in *Proc. INTERSPEECH*, 2012.

[53] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[54] Nicholas CV Pilkington, Heiga Zen, and Mark JF Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, 2011.

[55] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *IEEE SIGNAL PROCESSING LETTERS*, vol. 19, no. 12, pp. 914–917, 2012.

[56] Daisuke Saito, Shinji Watanabe, Atsushi Nakamura, and Nobuaki Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1784–1794, 2012.

[57] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, 2009.

[58] P Song, YQ Bao, L Zhao, and CR Zou, "Voice conversion using support vector regression," *Electronics letters*, vol. 47, no. 18, pp. 1045–1046, 2011.

[59] Elina Helander, Hanna Silén, Tuomas Virtanen, and Moncef Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[60] Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "Conditional restricted boltzmann machine for voice conversion," in *the first IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

[61] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *the 8th ISCA Speech Synthesis Workshop*, 2013.

[62] David Sundermann and Hermann Ney, "VTLN-based voice conversion," in *Proc. the 3rd IEEE International Symposium on Signal Processing and Information Technology*, 2003.

[63] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[64] Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[65] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[66] Seyed Hamidreza Mohammadi and Alexander Kain, "Transmutative voice conversion," in *Proc. ICASSP*, 2013.

[67] David Sundermann, Harald Hoge, Antonio Bonafonte, Hermann Ney, Alan Black, and Shri Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. ICASSP*, 2006.

[68] Thierry Dutoit, A Holzapfel, Matthieu Jottrand, Alexis Moinet, Javier Perez, and Y Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. ICASSP*, 2007.

[69] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. INTERSPEECH*, 2013.

[70] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012, pp. 1–5.

[71] Zhizheng Wu, Anthony Larcher, Kong Aik Lee, Eng Siong Chng, Tomi Kinnunen, and Haizhou Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints," in *Proc. INTERSPEECH*, 2013.

[72] Jean-François Bonastre, Driss Matrouf, and Corinne Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. INTERSPEECH*, 2007.

[73] Federico Alegre, Ravichander Vipperla, Nicholas Evans, and Benoıt Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012.

[74] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. ICASSP*, 2012.

[75] Zvi Kons and Hagai Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems," in *Proc. INTERSPEECH*, 2013.

[76] Hagai Aronowitz, Ron Hoory, Jason Pelecanos, and David Nahamoo, "New developments in voice biometrics for user authentication," in *Proc. INTERSPEECH*, 2011.

[77] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "RSR2015: Database for text-dependent speaker verification using multiple pass-phrases.," in *Proc. INTERSPEECH*, 2012.

[78] Driss Matrouf, J-F Bonastre, and Corinne Fredouille, “Effect of speech transformation on impostor acceptance,” in *Proc. ICASSP*, 2006.
[79] Zhizheng Wu, Eng Siong Chng, and Haizhou Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” in *Proc. INTERSPEECH*, 2012.
[80] Federico Alegre, Ravichander Vipperla, Nicholas Evans, et al., “Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals,” in *Proc. INTERSPEECH*, 2012.
[81] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, “Synthetic speech detection using temporal modulation feature,” in *Proc. ICASSP*, 2013.