SIP (2014), vol. 3, e17, page 1 of 16 © The Authors, 2014.

doi:10.1017/ATSIP.2014.17

OVERVIEW PAPER

Voice conversion versus speaker verification: an overview

ZHIZHENG WU[1,2,†] AND HAIZHOU LI[1,3]

A speaker verification system automatically accepts or rejects a claimed identity of a speaker based on a speech sample. Recently, a major progress was made in speaker verification which leads to mass market adoption, such as in smartphone and in online commerce for user authentication. A major concern when deploying speaker verification technology is whether a system is robust against spoofing attacks. Speaker verification studies provided us a good 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 attacks; therefore, presents a threat to speaker verification systems. In this paper, we will briefly introduce the fundamentals of voice conversion and speaker verification technologies. We then give an overview of recent 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.

Keywords: Speaker verification, Voice conversion, Spoofing attack, Anti-spoofing, Countermeasure, Security

Received 29 May 2014; Revised 1 December 2014; Accepted 1 December 2014

I. INTRODUCTION

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

Just like any other means of biometrics, an automatic speaker verification (ASV) system is not only expected to be 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 a sensed signal. At the sensor level, an adversary, that we call an impostor, could deceive the system by impersonating a target speaker at the microphone, or replace the acquired voice signal by a synthetically generated signal or imitated voice at the transmission time. In general, spoofing attack is to use a falsifying speech signal as system input for feature extraction and verification; therefore, presenting a threat to speaker verification systems. In this paper, an *impostor* means a *zero-effort impostor* who spoofs a system without relying on any technology, while we call a *non-zero-effort impostor* as *attacker*, who uses voice conversion or other technique to mimic the target speaker.

As digital recording has become widely accessible, *replay attack* is the simplest way to deceive 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 speaker verification (TI-SV) systems which do not impose constraints on linguistic content [3, 4]. However, if the replayed content is different from the specific pass-phrase required by a text-dependent speaker verification (TD-SV) system, it does not pose a threat unless the attack is able to acquire the target speaker's voice for that specific pass-phrase as assumed in [5].

Aside from replay attack, *human voice mimicking* or *impersonation* has also received considerable attention [6–8]. As impersonation requires special skills, it is difficult to judge its effectiveness as a general 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 [6]. Professional voice mimics, often voice actors, tend to mimic prosody, accent, pronunciation, lexicon, and other high-level speaker traits, rather

[1]School of Computer Engineering, Nanyang Technological University, Singapore 639798
[2]Temasek Lab@NTU, Nanyang Technological University, Singapore
[3]Institute for Infocomm Research, Singapore 138632

Corresponding author:
Zhizheng Wu
Email: zhizheng.wu@ed.ac.uk
[†]Present address: Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.

than 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. Owing to the rapid development of *unit selection* [9], *statistical parametric* [10], and *hybrid* [11] methods, speech synthesis systems are now able to generate speech with acceptable quality as well as voice characteristics of a given target speaker, such as spectral cues. In early studies [12–14], vulnerability of text-prompted *hidden Markov model* (HMM)-based speaker verification was examined using a small database of 10 speakers. More recently, [15] used a flexible adapted HMM-based speech synthesis system to spoof TI-SV systems on a corpus of around 300 speakers. Although HMM-based synthesis poses a threat especially to TD-SV system, usually hours of training speech are needed to train a speech synthesis system of reasonable quality. Even an adapted HMM-based speech synthesis system requires a significant amount of speakers' data to train an average voice model for target speaker adaptation [16]. Therefore, it is not as straightforward as people think to use HMM-based speech synthesis to impersonate someone's voice.

Different from replay attack, human voice mimicking and speech synthesis, *voice conversion* transforms one speaker's (source) voice to sound like that of another speaker (target) without changing the language content. 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 sound like 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 cannot distinguish easily, it presents a genuine threat to both text-dependent and TI-SV systems.

In summary, human voice can be seen to have three attributes, language content, spectral pattern, and 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. Human voice mimicking typically modifies the prosodic patterns while voice conversion modifies both spectral and prosodic patterns. As it is more reliable to characterize speakers by their spectral cues [17], 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 voice conversion spoofing attacks, and review the most recent research works on voice conversion, speaker verification, spoofing attack, and anti-spoofing attack techniques. A general review on spoofing and anti-spoofing for speaker verification can be found in [18].

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

II. VOICE CONVERSION TECHNIQUES

Human voice conveys not only language content but also speaker individuality. From the perspective of speech perception, speaker individuality is characterized at three different levels: segmental, supra-segmental, and linguistic information. The segmental information relates to the short-term feature representations, such as spectrum and instantaneous fundamental frequency (F_0). The supra-segmental information describes prosodic features such as duration, tone, stress, rhythm over longer stretches of speech than phonetic units. It is more related to the signal but spanning a longer time than the segmental information. The linguistic information is encoded and expressed through lexical words in a message. Since each speaker has his/her own lexical preference, the choice of words and sentence structures, the same linguistic information can be conveyed by different people in different ways.

Voice conversion technology is to deal with the segmental and supra-segmental information while keeping the language content unchanged. In particular, the objective of voice conversion is to modify one speaker's voice (source) to sound like another speaker (target) without changing the language content. Mathematically, voice conversion is a process to learn a conversion function $\mathcal{F}(\cdot)$ between source speech $\mathbf{Y}$ and target speech $\mathbf{X}$, and to apply this conversion function to a source speech signal $\mathbf{Y}$ at runtime in order to generate a converted speech signal $\hat{\mathbf{X}}$. This process is formulated as follows:

$$\hat{\mathbf{X}} = \mathcal{F}(\mathbf{Y}). \tag{1}$$

Figure 1 presents a typical voice conversion framework, which consists of off-line training and runtime conversion processes. During off-line training, features, which characterize the speaker individuality, in the form of parameter vectors are first extracted from source and target speech signals. Then, each source feature is paired up with one target feature, which is called frame alignment, to establish the source–target correspondence. The frame alignment is usually achieved through dynamic time warping for parallel data [19], or through some advanced frame alignment techniques for non-parallel data [20]. Finally, a conversion function is estimated from the source–target feature pairs.

At runtime, the conversion function is employed to the features extracted from source speech, and then the converted feature vector sequence is passed to a synthesis filter to reconstruct an audible speech signal. Next we discuss feature extraction and the estimation of conversion function in a greater detail.

A) Feature extraction

In voice conversion, we consider two levels of features, namely short-term spectral and prosodic features, that correspond to the segmental and supra-segmental information.

Short-term spectral features are to represent the spectral attributes that relate to voice timbre. Mel-cepstral

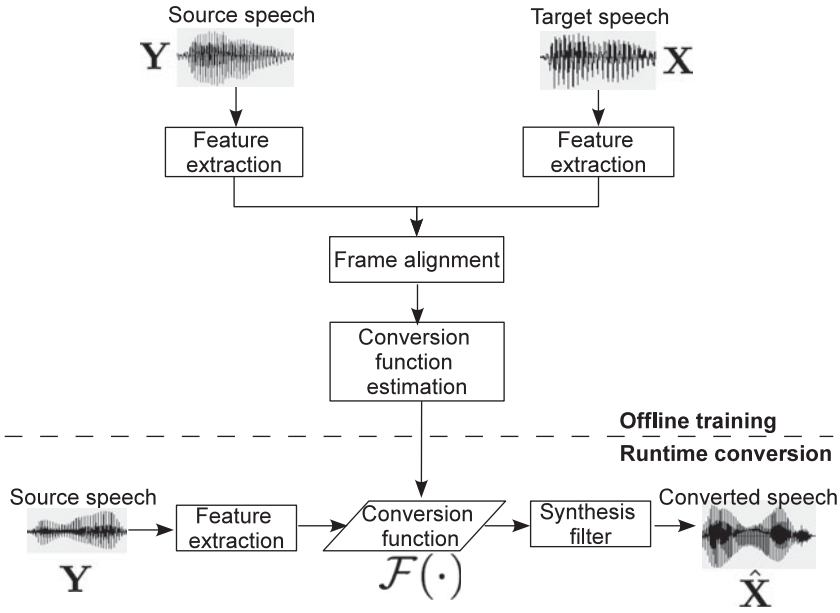

Fig. 1. Diagram of a typical voice conversion system.

coefficients (MCCs), linear predictive cepstral coefficients (LPCCs), and line spectrum frequency (LSF) are the popular short-term spectral features to represent the spectral envelope for voice conversion. The dynamic features, such as delta and delta–delta features, can also be employed to capture the speech dynamics to generate converted speech of better quality. Formant feature is another kind of short-term feature representation to describe the vocal tract, and has been employed in some voice conversion systems.

Prosodic features also include significant speaker individualities. Intonation, duration, and intensity are typical prosodic features. Intonation represents the fundamental frequencies contour over a longer time, and describes the tones of syllables as well as the accent of a speaker.

B) Conversion function

Spectral mapping and prosodic conversion map the segmental and supra-segmental information, respectively, from one speaker to another. We next discuss these two forms of conversion functions.

1. Spectral mapping

The spectral mapping methods can be roughly grouped into three categories: statistical, frequency warping, and unit-selection methods.

In the statistical methods, the relationship between source and target features is established through parametric models. They are used to implement the conversion function to map source feature into target feature space. Vector quantization (VQ) is a simple and straightforward mapping method, which was proposed in [21]. This method implements a codebook from the paired source–target features. The codebook is used to find the corresponding target vector for each source feature vector. Some statistical models have been proposed to improve the VQ method. The Gaussian mixture model (GMM) [22–24], partial least-squares regression [25], and trajectory HMM [26] are good examples that assume a linear relationship between the source and the target features. Assuming a non-linear relationship between the source and target speech features, researchers studied another group of methods, such as artificial neural network [27–31], support vector regression [32], and kernel partial least-squares regression [33].

In the statistical methods, the conversion function is formulated from the parametric representations of the spectrum without following a physical principle. Therefore, the statistical averaging effect, which reflects the central tendency of speech features, could introduce oversmoothing [24, 34, 35]. Frequency warping methods take the physical principles into consideration and aim to warp the frequency axis of the amplitude spectrum to the source speaker to match that of the target speaker [36–41]. In this way, the frequency warping methods are able to keep more spectral details and produce high-quality converted speech. The basic frequency warping methods only consider shifting the frequency axis without taking the amplitude into consideration. To bridge this gap, an amplitude scaling technique was proposed in [39] to enhance the conversion performance. 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 [40].

Generally speaking, the statistical parametric and frequency warping methods attempt to modify the speaker characteristics. Unlike these methods, unit-selection methods utilize original target speaker's feature vectors to construct the converted [42–44]. This idea is inspired by the unit-selection for speech synthesis [9]. In voice conversion, as training data are limited, the basic unit generally spans only one [42, 43] or several frames [44].

2. Prosodic conversion

Prosodic conversion relates to the prosodic features, such as fundamental frequency, intonation, and duration. The most simple and common approach is to normalize the mean and variance of the (log-)F_0 distribution of the source speaker to those of the target speaker. This approach operates on instantaneous F_0 value and only changes the global level of the F_0 as well as the F_0 range. However, the target voice takes the same duration and intonation pattern as the source voice.

Some attempts have been made to extend the mean-variance normalization (MVN) approach, such as higher-order polynomial [45], piecewise linear transformation [46], and GMM-based mapping [47, 48]. These approaches also operate on the instantaneous F_0, and work well if the source and target speakers have "similar" intonation patterns. Instead of operating on instantaneous F_0, more advance methods were proposed in [45, 47, 49, 50] to convert intonation patterns directly at syllable level or even longer segments. These methods usually require manually labelling the intonation boundaries.

In addition to the F_0/intonation conversion, duration conversion was proposed in [51–54]. Duration is related to the rhythm and tempo in a speech signal, and is one of the important factors to describe speaker individuality. In [51], duration-embedded Bi-HMMs were proposed to convert spectral attributes and duration simultaneously. Bi-HMMs mean parallel source-target HMMs capturing the source and target features. In [52], a probabilistic model was proposed to deal with two different length utterances, where the frame alignment between source and target feature sequences was represented through hidden variables. A similar idea was presented in [54] to simultaneously convert duration and spectrum. In [53], the syllable-level duration was converted through maximum-likelihood linear regression (MLLR), and relaxed the requirement of parallel data.

3. Summary

In general, spectral/prosodic mapping techniques are to match the spectral/prosodic attributes of the target speaker given the source speaker's spectral/prosodic features. As discussed above, a large number of approaches have been proposed aiming to improve the quality of voice conversion. Here we are more interested in the effectiveness of voice conversion methods for spoofing attacks.

From the perspective of spectral mapping, both statistical and frequency warping methods are flexible when the training data are limited, while unit-selection methods are expected to achieve better performance when sufficient data, for example 30 min speech, are available. In the statistical methods, the maximum-likelihood Gaussian mixture model (ML-GMM) with dynamic feature constraint method [24] and the dynamic kernel partial least-squares method (DKPLS) [33] are two popular methods that achieve stable performance with different amount of training data. In particular, the ML-GMM method is a well-established baseline method in the voice conversion research. In the frequency warping methods, the weighted frequency warping with amplitude scaling (WFW-AS) has been reported to achieve comparable performance to ML-GMM in terms of speaker similarity [39]. Hence, ML-GMM, DKPLS, and WFW-AS could be good choices to simulate voice conversion spoofing attacks when the training data are limited, although not all of them have been applied to spoofing attacks.

In prosodic conversion, the conversion of intonation pattern requires manually labeling of intonation boundaries and patterns as well as a large amount of training data. The most practical way is to do mean and variance normalization on F_0 values.

III. 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 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$ [17]. This verification process is illustrated in Fig. 2 and is formulated as a hypothesis test:

$$\Lambda(\mathbf{X}) = \frac{p(\mathbf{X}|\lambda_H)}{p(\mathbf{X}|\lambda_{\widetilde{H}})}, \tag{2}$$

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

In this section, we will briefly describe the state-of-the-art techniques in speaker verification systems that relate to voice conversion spoofing attacks. The techniques include feature extraction to obtain representations for the speech sample $\mathbf{X}$ and the speaker modeling for the models λ_H and $\lambda_{\widetilde{H}}$. More general overviews or tutorials on speaker verification can be found in [17, 55–60].

A) Feature extraction

In Section II, we consider three levels of information, namely segmental, supra-segmental, and linguistic information that describe speaker individuality, correspondingly, there are three level of features to characterize the individuality of speakers: spectral, prosodic, and high-level features [17]. All the three levels of features are generally used in ASV.

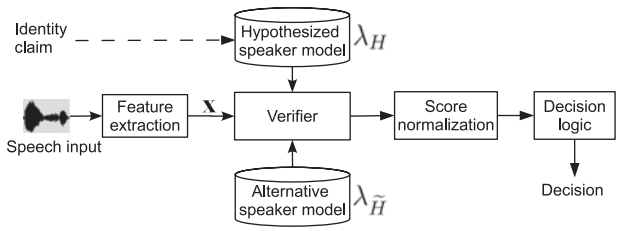

Fig. 2. Diagram of a speaker verification system.

As speech signals are not stationary, shifting windows are generally applied to divide speech signals into short-term overlapping segments with about 20–30 ms. The short-term spectral features, such as mel-frequency cepstral coefficient (MFCC), LPCC, LSF, and perceptual linear prediction (PLP), are generally extracted from the short-term speech segments. The short-term dynamic features, such as delta and delta–delta coefficient of these short-term features, are usually computed to take the speech dynamics into account. Different from short-term spectral features, temporal modulation features are another types of spectral features that extracted from multiple consecutive spectral segments [61, 62].

Prosodic features, such as intonation, intensity, and duration, are corresponding to the supra-segmental information. Prosodic features have been adopted in speaker verification in [63–66]. Although these features are usually more robust than short-term features in face of channel variations, the extraction of these features are also affected by noise. For example, the fundamental frequency cannot be well estimated in noise environment, and as such the accuracy of intonation pattern will be affected.

High-level features, such as phoneme, pronunciation, and the choice of words in conversation, are more related to lexical information. These features are more robust against noise comparing with other levels of features. However, they rely on other sophisticated techniques such automatic speech recognition, thus they are more difficult to use.

Recognizing the high effectiveness of short-term spectral features, most of the speaker verification systems adopt short-term spectral features in the implementation.

B) Speaker modeling

There are two kinds of speaker verification systems: TI-SV and TD-SV systems. TD-SV assumes cooperative speakers and requires the speaker to speak fixed or spontaneously prompted utterances, whereas TI-SV allows the speaker to speak freely during both enrolment and verification. Both TI-SV and TD-SV systems share the feature extraction techniques, while differ in the speaker modeling.

1. Text-independent modeling

The modeling techniques for TI-SV can be grouped into three categories: generative, discriminative, and fusion models. Generative models focus on modeling the feature's distribution of a target speaker. The GMM [67], joint factor analysis (JFA) [68, 69], and probabilistic linear discriminant analysis (PLDA) [70] are typical generative models. GMM has been used intensively to model the feature distributions, and GMM with the universal background model (UBM) is the classic method in building speaker verification systems [67]. In the GMM–UBM method, speech samples from a large number of non-target speakers are first employed to build a speaker-independent UBM, and then the target speaker's samples are adopted to adapt the UBM to estimate a speaker-dependent GMM. During runtime, the target GMM and the UBM are used as hypothesized speaker model λ_H and the alternative speaker model $\lambda_{\tilde{H}}$, respectively.

JFA and PLDA, a latent variable model, are more advanced generative models, explicitly model the channel and speaker variabilities jointly. The JFA works within the GMM mean supervector space, whereas PLDA models the channel and speaker variabilities within i-vector space. An i-vector is a low-dimensional set of factors to represent speaker and channel information via factor loadings, also called total variability [71]. In both JFA and PLDA, a large number of additional data are required for estimating the speaker and channel variabilities, or the total variability.

Unlike generative models, discriminative models do not attempt to model the feature distributions, but to focus on the difference between the hypothesized speaker model and the alternative speaker model. Support vector machine (SVM) is a type of discriminative model that can be used together with the GMM–UBM or the i-vector framework. In [72], GMM mean supervectors are used as features to estimate an SVM classifier, and in [73], SVM is combined with the i-vector framework. Other SVM-based approaches were also proposed, such as SVM with score-space kernels [74]. In the context of SVM modeling, nuisance attribute projection (NAP) [75, 76] and within-class covariance normalization (WCCN) [77] techniques are proposed for channel compensation. Other discriminative models such as neural networks [78–85] have also been employed in speaker verification.

The fusion approach is trying to fuse multiple subsystems into one to benefit from multiple "experts". In the generative and discriminative models, they attempt to build an individual system, and in practice it is not enough to build just one single strong system. Such a fusion model assumes that individual systems are able to capture different aspects of a speech signal, and provide complimentary information for each other. Each individual system can involve different kinds of features or different level of features, and can also employ different modeling techniques. Although fusion usually takes place at the score level across subsystems [86–88], there are also ways to fuse the features or speaker models [88].

2. Text-dependent modeling

Different from the text-independent speaker modeling, text-dependent systems not only model the feature's distribution, but also model the linguistic information such as phonetic and prosodic patterns. For example, text-dependent systems use HMMs and other techniques that are developed in automatic speech recognition to capture the supra-segmental and linguistic features in the pass-phrases. In terms of decision strategy, text-dependent systems share the similar system architecture with that of text-independent systems (see Fig. 2). More on text-dependent modeling or classifiers can be found in [60].

C) Vulnerability of speaker verification to voice conversion

As discussed above, a speaker verification system makes a decision based on the feature distributions through speaker modeling. The feature extraction and speaker modeling modules are hence the two most important components. Accordingly, there are two classes of weak links, one in feature extraction and the other in speaker modeling.

From the perspective of feature representation, it is known to the public that speaker verification systems use spectral, prosodic, and linguistic features. Thus, speaker verification systems may be vulnerable to the attackers that can manage to mimic those features. On the other hand, voice conversion can modify or mimic all the three levels of features that are also used in speaker verification. Given a sequence of features $\mathbf{Y}$ from an attacker, voice conversion technology can project the attacker's features to the target speaker's feature space through the mapping function $\hat{\mathbf{X}} = \mathcal{F}(\mathbf{Y})$, and in this way, the speaker verification systems can be deceived by the generated target features $\hat{\mathbf{X}}$.

The spectral and prosodic features are popular features used in speaker verification. In particular, due to the simplify and robust performance, the spectral features are widely used in practical implementation. As discussed in Section A, MFCCs, LPCCs, and LSFs are the popular features to describe the spectral attributes, while F_0, intensity, duration, and intonation are shared by a large range of speaker verification systems to represent prosodic attributes. On the other hand, those spectral and prosodic features are also involved in voice conversion. Therefore, knowing how spectral or prosodic features are used in a speaker verification system, one is able to devise a spectral prosodic mapping that generates spectral or prosodic features to deceive a speaker verification system.

There are also weak links from the linguistic or high-level feature aspects. In the TD-SV case, it is possible for the attacker to obtain the exact pass-phrase information in advance, while for the TI-SV case, the attacker can either familiarize the choice of words and speaker style of the target speaker in advance or speak freely, as TI-SV systems do not have any constraint in the language content for verification.

From the perspective of speaker modeling, most of the systems use a GMM as the basis to model feature distributions. Such an implementation ignores the temporal structure of speech, which also reflects the speaker individuality. On the other hand, voice conversion systems are good at performing frame by frame conversions. In this way, the loss of temporal structure modeling in speaker verification is a weak link to spoofing attacks. Studies have shown that HMM-based speaker verification systems that capture the temporal structure are more resilient than those without temporal constraint in the face of voice conversion spoofing attacks [89]. But we need to note that the latest voice conversion systems, such as duration embedded HMM [51] and trajectory HMM [26]-based systems, are designed to transfer the temporal structure of speech from source to target speaker. Hence, whether temporal modeling techniques can provide some protection to voice conversion spoofing remains an open question.

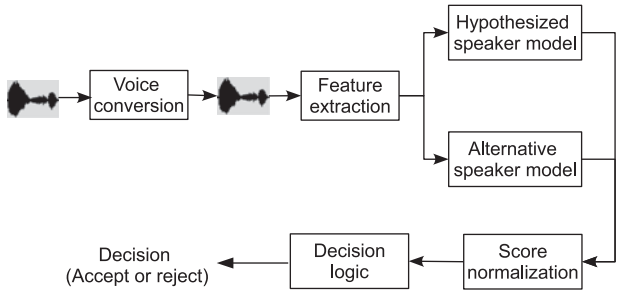

Fig. 3. Illustration of a voice conversion spoofing process, in which an attacker's voice is modified by a voice conversion system and then passed to a speaker verification system for verification.

IV. SPOOFING ATTACK STUDIES

With voice conversion as in equation (2), we modify the source speech $\mathbf{Y}$ to sound like that of a target speaker X, and this presents a threat to speaker verification systems. Figure 3 illustrates a general voice conversion spoofing attack process.

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

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

A) Evaluation metrics

In speaker verification, the decision of a test sample or a trial belongs to one of the four categories as presented in Table 1. If the speaker identity of the test sample matches that of the hypothesized model or the claimed speaker, then we call it a genuine test; otherwise, an impostor test. If a genuine test is rejected as an impostor, then it is a miss detection or false rejection decision; similarly, if an impostor test is accepted as a genuine speaker, then it is a false alarm or false acceptance. The equal error rate (EER) is a common evaluation measure to balance the false acceptance rate (FAR) and the false rejection rate (FRR). The EER is one of the popular criterion to optimize speaker verification systems.

In the voice conversion spoofing scenario, an attacker attempts to use voice conversion technology to modify his/her voice to sound like a target genuine speaker in order to deceive a speaker verification system. The purpose of voice conversion spoofing is to fool speaker verification systems as a result to increase FARs. Thus, the FAR is a good vulnerability indicators of speaker verification systems under voice conversion attacks. In the experiments, if the genuine trials are kept the same, the increase in the FARs will result in an increase in the EERs. Therefore, it

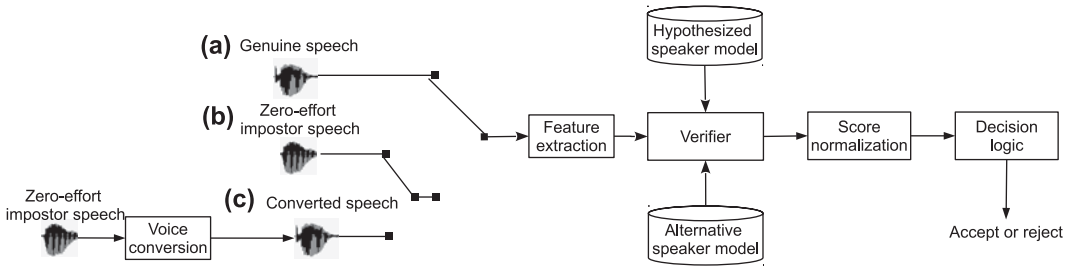

Fig. 4. Illustration of the vulnerability evaluation framework used in the past studies. The figure involves three kinds of trials: (a) genuine speech; (b) impostor speech; and (c) converted speech. (c) is a converted version of (b). (a) and (b) make a standard speaker verification test, whereas (a) and (c) make a spoofing test.

Table 1. Four categories of trial decisions in speaker verification.

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

Table 2. Subset of NIST SRE 2006 core task in the spoofing attack experiments [90] (VC = voice conversion).

	Baseline test	Spoofing test
Unique speakers	504	504
Genuine trials	3978	3978
Impostor trials	2782	0
Spoofed trials (impostor trials via VC)	0	2782

is easy to understand that the majority of the past studies use both EER and FAR as evaluation metrics to measure the vulnerability of speaker verification system against voice conversion spoofing.

B) Database design

In the past studies, several different datasets have been used to provide an objective assessment of system performance under voice conversion attacks. There are some similarities in the design of datasets and the experimental protocols among them. In this paper, we use the dataset in [90] as a case study to show the common protocol in designing the spoofing dataset. This dataset is based on the National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE) 2006 core task, namely *1conv4w-1conv4w*.

The common framework used in the majority of the past studies is represented in Fig. 4. Different from a standard speaker verification experiments, the spoofing attack experiments generally have three kinds of trials: genuine, zero-effort impostor, and spoofing trials.

The genuine and zero-effort impostor trials are directly selected from the original core task, *1conv4w-1conv4w*. The training data for each target speaker model are also a subset of the core task. To generate the spoofing trials, the attackers and their corresponding target genuine speakers are first selected. Then, the data from the *3conv4w* and *8conv4w* training sections in the NIST SRE 2006 database are employed to estimate the conversion function for each impostor–target speaker pair. Finally, each zero-effort impostor trial is passed through the conversion function to generate its corresponding spoofed trial. It is noted that the number of spoofed trials is exactly the same as that of the zero-effort impostor trials, and the genuine trials are kept unchanged as in the original test.

In the experiment, the genuine trials and the zero-effort impostor trials are mixed as a baseline test, while the same genuine trials and the spoofed trials are mixed as a spoofing test. In this way, the baseline results in terms of EERs and FARs are comparable with the spoofing results; furthermore, with such a comparison, the vulnerability of speaker verification under voice conversion attacks can be assessed and predicted. The statistics of trial used in the case study is presented in Table 2. This setup may be different from an actual real-world scenario where live impostor trials and spoofed trials are mixed together, but it allows us to conduct an analytical study under an extreme adverse condition.

C) Experiments

A number of studies have been conducted to evaluate the vulnerability of speaker verification systems under voice conversion attacks. The earlier work involves GMM–UBM speaker verification systems. The vulnerability of a GMM–UBM speaker verification system was assessed in [95] for the first time. The YOHO corpus consisting of 138 speakers was employed to design the spoofing dataset. The experiments showed that the baseline FAR increased from 1.45 to 86.1% as a result of voice conversion attack. In [96], the vulnerability of a GMM–UBM speaker verification system was evaluated using the NIST SRE 2004 dataset. The experimental results showed that the baseline EER and FAR increased from both 16 to 26% and over 40%, respectively, as a result of voice conversion spoofing.

The work in [97] evaluated the vulnerability of a GMM–UBM system under voice conversion attack, and the spoofing attack was simulated by a Gaussian-dependent filtering voice conversion approach, which shift the spectral shape of the attacker toward that of the target genuine speaker. The experimental results reported on the NIST SRE 2005 database showed that the baseline EER and FAR increased from both 8% to over 60 and 100%, respectively. Note the

Table 3. Summary of voice conversion spoofing attack studies (TI, text-independent recognizer; TD, text-dependent).

					Spoofing	
Conversion method	Database	TI or TD	System	Baseline EER (%)	EER (%)	FAR (%)
Filtering [91]	NIST SRE 2005	TI	GMM–UBM	8.54	35.41	N. A.
Filtering [91]	NIST SRE 2006	TI	GMM–UBM	6.61	28.07	N. A.
Filtering [92]	NIST SRE 2005	TI	GMM–UBM	8.50	32.60	N. A.
Filtering [92]	NIST SRE 2005	TI	JFA	4.80	24.80	N. A.
JD-GMM [93]	NIST SRE 2006	TI	GMM–UBM	7.63	24.99	N. A.
JD-GMM [93]	NIST SRE 2006	TI	VQ–UBM	7.56	22.62	N. A.
JD-GMM [93]	NIST SRE 2006	TI	GMM–SVM	3.74	12.58	41.54
JD-GMM [93]	NIST SRE 2006	TI	JFA	3.24	7.61	17.33
Unit-selection [90]	NIST SRE 2006	TI	JFA	3.24	11.58	32.54
JD-GMM [90]	NIST SRE 2006	TI	PLDA	2.99	6.77	19.29
Unit-selection [90]	NIST SRE 2006	TI	PLDA	2.99	11.18	41.25
Filtering [89]	WF corpus [94]	TI	I-vector	1.60	8.80	29.00
Filtering [89]	WF corpus [94]	TI	GMM–NAP	1.10	3.40	38.00
Filtering [89]	WF corpus [94]	TD	HMM–NAP	1.00	2.90	36.00

full knowledge, for example feature extraction and speaker modeling, of the speaker verification system was assumed in the experiments. Using the same voice conversion method, the authors in [91] evaluated the GMM–UBM verification system on both NIST SRE 2005 and NIST SRE 2006 databases. The EERs increased from 8.54 and 6.61% to 35.41 and 28.07% on NIST SRE 2005 and 2006 databases, respectively. Different from [97], the work in [91] did not assume any prior information of the speaker verification system.

In addition to the GMM–UBM systems, in [93] and [90], the vulnerabilities of six state-of-the-art speaker verification system were assessed under the same voice conversion attack. The spoofing attack was simulated by the joint-density Gaussian mixture model (JD-GMM) voice conversion method. The experimental results showed that the EERs increased more than two times over those of the baselines for all the text-independent systems. The EER and FAR of the JFA system increased from both 3.24% to 7.61 and 17.33%, respectively, and the EER and FAR of the most robust PLDA system increased from both 2.99% to 11.18 and 41.25%, respectively. Such increase in EER and FAR is due to the shift of classifier score as a result of voice conversion attack, as presented in Fig. 6. It is clearly observed that after voice conversion attack, the impostor trials' score distribution moves toward that of the genuine trials.

Still in the context of text-independent ASV, other work relevant to voice conversion includes attacks referred to as artificial signals. It was noted in [92] and [98] that certain short intervals of converted speech yielded extremely high scores or likelihoods. Such intervals are not representative of intelligible speech but are nonetheless effective in overcoming ASV systems which lack any form of speech quality assessment. Artificial signals optimized with a genetic algorithm were shown to provoke increases in EER from 8.5% to almost 80% for a GMM–UBM system and from 4.8% to almost 65% for a factor analysis (FA) system.

The work in [89] examined the vulnerability of several state-of-the-art text-dependent systems, namely, i-vector, GMM–NAP, and HMM–NAP systems. Among the three systems, HMM–NAP employed a speaker-independent HMM instead of a GMM to capture temporal information. The results showed that voice conversion provoked increases in the EERs and FARs of all the three systems. Specifically, the FAR of the most robust HMM–NAP system increased from both 1–36%.

Table 3 presents a summary of spoofing studies described above. Even though some approaches to voice conversion produce speech with clearly audible artifacts [24, 34, 99], Table 3 shows that all provoke significant increases in the FAR across a wide variety of different ASV systems. Figure 5 presents a comparison of spectrograms and formant tract of impostor speech, impostor speech after voice conversion, and genuine speech. It shows that as a result of voice conversion, the impostor speech is shifted toward that of the genuine speaker. Such speech or feature shifting explains the score shifting and FAR increasing as a result of voice conversion spoofing.

V. ANTI-SPOOFING ATTACK STUDIES

As shown in Section IV, the EER performance of most speaker verification systems degrades considerably under voice conversion spoofing attacks. It is hence necessary to develop anti-spoofing measures to enhance the security of speaker verification systems. The key to develop a spoofing-proof speaker verification system is two folds. One is to characterize speakers with unique features and models that voice conversion techniques cannot reproduce easily [26, 51]. The other one is to detect artifacts that come along with voice conversion [100], that is to design a countermeasure for anti-spoofing. In this section, we review the past effort on designing countermeasures in the form of converted speech detectors for speaker verification systems in the face of voice conversion spoofing.

We have seen successful techniques that detect artifacts introduced during the voice conversion or synthesis

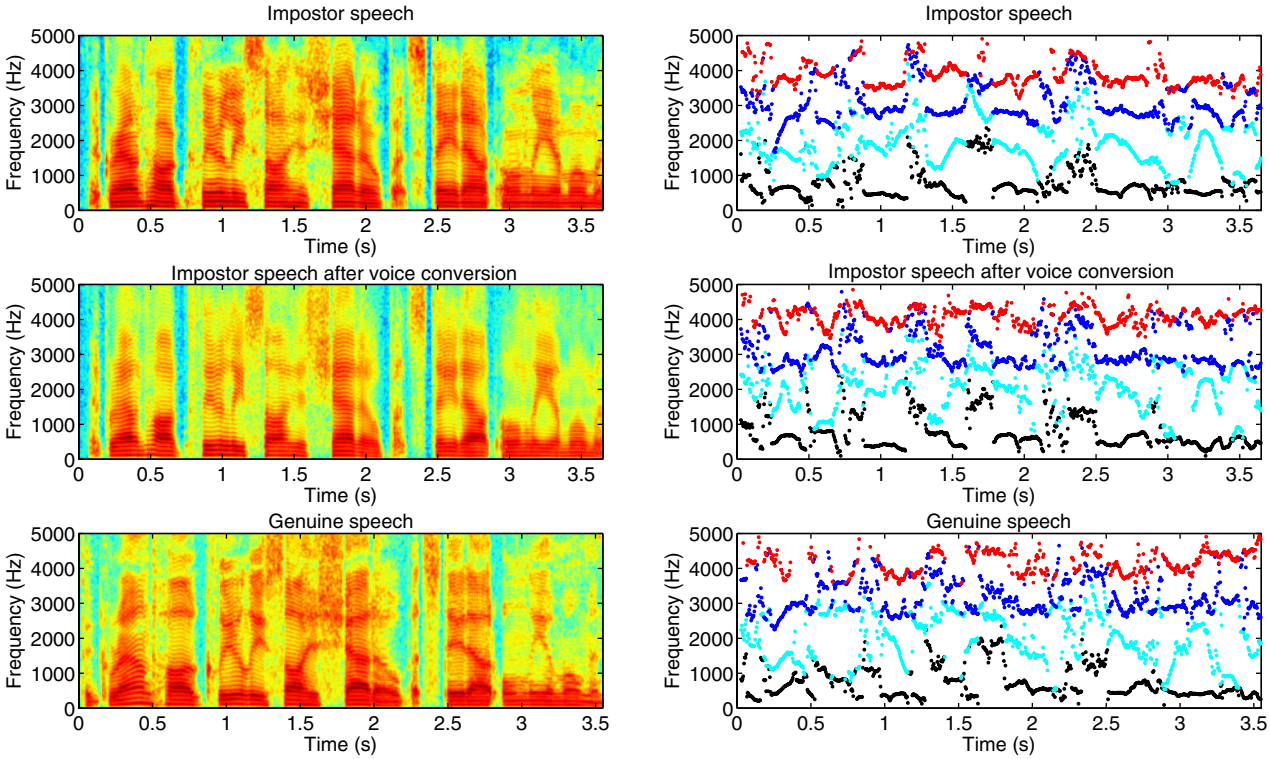

Fig. 5. An illustration of voice conversion spoofing. An attacker attempts to use voice conversion to shift his/her voice (top) toward the target genuine speaker's voice (bottom), and generates a modified voice (middle). From the spectrograms (left column) and the formant tracks (right column), it shows that after voice conversion, the impostor's speech is much closer to the target genuine speaker's speech. This explains the phenomenon of score shifting as a result of voice conversion spoofing.

process. In [100], Cosine normalized phase (cos-phase) and modified group delay phase (MGD-phase) features were proposed to detect converted speech. They are motivated by the fact that most vocoders use minimum-phase rather than original phase to reconstruct a speech signal. We note that most vocoders assume that human auditory is not sensitive to the phase information, and hence the original phase information is discarded when synthesizing speech signals. As MGD-phase feature not only contains phase information, but also the magnitude information, it is sensitive to vocoder outputs. Figure 7 presents an example of the MGD spectrogram. It is clearly observed that the MGD spectrograms between the original and converted speech signals are different. The experiments on NIST SRE 2006 database were reported to obtain a detection EER of 5.95 and 2.35% using cos-phase and MGD-phase, respectively, confirming the effectiveness of the phase-based detectors.

The MGD-phase-based detector was integrated with speaker verification systems, in particular GMM–JFA and i-vector PLDA systems, in [90] for anti-spoofing. Figure 8 presents an example of incorporating a converted speech detector as an explicit countermeasure against spoofing attacks. Two GMMs were trained from the natural and the converted speech, respectively, and the natural or converted speech decision was made based on log-likelihood ratio. The experimental results reported on the NIST SRE 2006 confirmed the effectiveness of the MGD-phase-based detector. The converted speech detector can reduce the FARs from 17.36 and 19.29% to both 0.0% for GMM–JFA and PLDA systems, respectively, under GMM-based voice

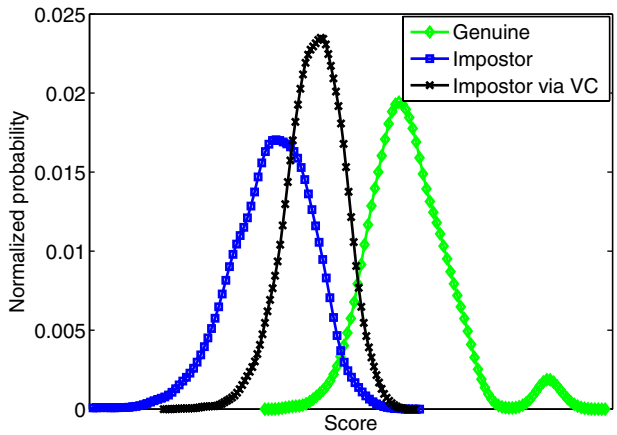

Fig. 6. Score distribution before and after voice conversion attack.

conversion spoofing, and reduce the FARs from 31.54 and 41.25% to 1.64 and 1.71% for GMM–JFA and PLDA systems, respectively, in the face of unit-selection-based voice conversion attacks. Interestingly, such a detector works well in the face of spoofing attacks, and it also does not affect the speaker verification performance in the face of non-zero-effort spoofing or normal genuine tests.

In [98], a long-term dynamic feature, which was extracted at the utterance-level, was proposed to capture the utterance-level speech variation for detecting converted speech. The experimental results reported on the NIST SRE 2005 showed the effectiveness of such a long-term feature in distinguishing the converted or the so-called artificial

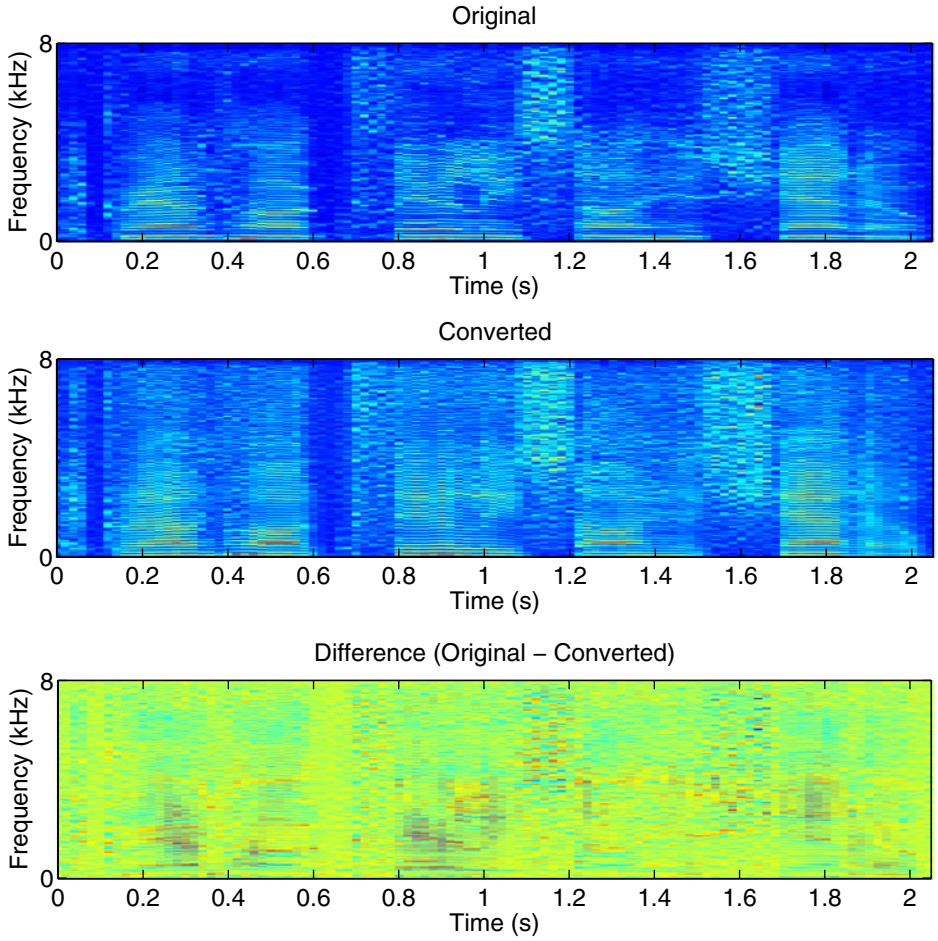

Fig. 7. An example of the MGD spectrogram. The MGD phase feature is extracted from such a spectrogram instead of a magnitude spectrogram. Top: MGD spectrogram of the original speech signal. Middle: MGD spectrogram of the corresponding converted speech signal. Bottom: the difference between the original and converted MGD spectrograms.

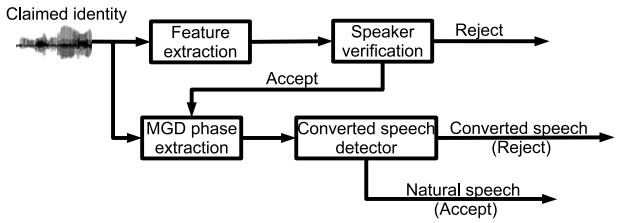

Fig. 8. Diagram of speaker verification with an anti-spoofing converted speech detector [90] (MGD = modified group delay).

speech from the natural human speech. More specifically, an EER of 0.0% was achieved in the converted speech detection task. It is true that speech variation becomes small if voice conversion systems suffer over-smoothing. However, the global variance (GV) enhancement as proposed in [24] is able to recover the speech variation for better speech quality. It would be interesting to re-evaluate the effectiveness of such a long-term dynamic feature on GV enhanced speech.

Currently, the analysis–synthesis techniques for extracting feature representation and reconstructing audible speech signals operate on the short-term feature level, for example 5–15 ms; hence some artifacts are introduced in the temporal domain. In the wake of such artifacts, temporal modulation features, magnitude and phase modulation features, were proposed in [101] to detect the converted speech which is generated through vocoding techniques. This work assumes that no specific voice conversion method is required to design the detection, only utilizing the copy-synthesis speech. A copy-synthesis speech is obtained using a speech analysis module to extract feature representations from a natural speech signal, and then passing these feature representations through a matched vocoder to reconstruct an audible speech signal. The experiments conducted on the Wall Street Journal (WSJ0+WSJ1) database showed that the modulation feature-based detection achieved an EER of 0.89% in the synthetic speech detection task, while the baseline MGD-phase feature gave an EER of 1.25%.

Figure 9 illustrates a way to extract modulation features. The spectrogram, which can be a power spectrogram or an

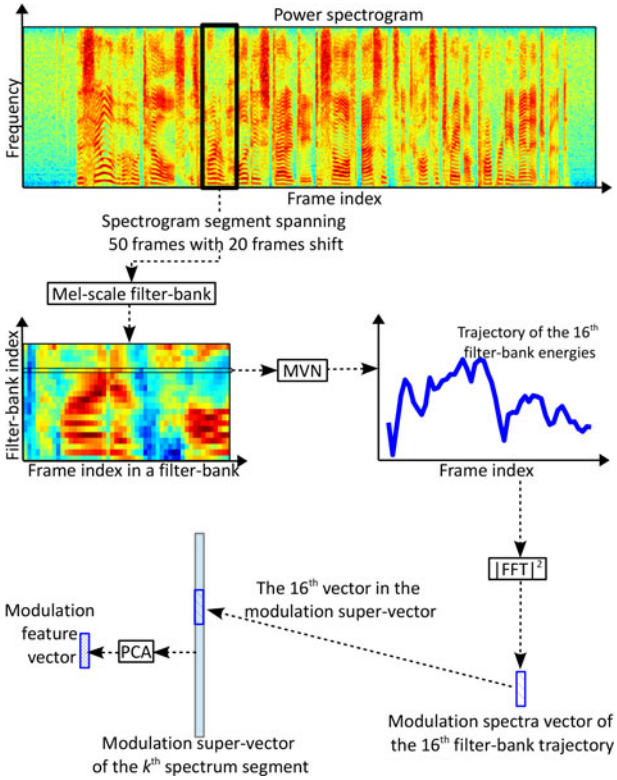

Fig. 9. Illustration of one way to extract modulation features from a spectrogram. The figure is adopted from [101].

MGD spectrogram, is first divided into overlapping short segments, for example a 50-frame segment with 20-frame shift. Then, filter-banks are applied to the spectrogram segments to obtain filter-bank coefficients. After that, the segment-level MVN is applied to the filter-bank trajectories to normalize the mean and variance to zero and unit, respectively. Next, fast Fourier transform (FFT) is adopted to transform the filter-bank trajectories into modulation spectra. The modulation spectrum from each filter-bank trajectory is stacked into a supervector, which undergoes principal component analysis (PCA) for dimensionality reduction. Finally, the low-dimensional compressed feature is used as the modulation vector. Meanwhile, modulation compensation is being investigated in speech synthesis for better quality [102]. Hence, the modulation feature-based detector might be countered if the modulation compensation techniques work well. Further attention is required to fully understand the effectiveness of modulation features in the context of more advanced synthesis techniques.

In [103], a local binary pattern (LBP) analyzed feature was proposed for anti-spoofing. The LBP analysis has been widely used in face recognition for texture analysis [104], and liveness detection [105]. The LBP feature is a kind of spectrotemporal feature, taking into account the local dynamics in the sequence of speech feature vectors. The experimental results reported on the male subset of the NIST SRE 2006 database showed that the LBP feature achieves an EER of 8% in the converted speech detection task. By integrating the LBP-based countermeasure with the FA speaker verification system, the FAR decreases from 54 to 4.3% in the face of voice conversion attacks. Note that the baseline performance is 1%.

The LBP-based countermeasure proposed in [103] was extended to one-class classifier in [106]. In the work, the LBP features were extracted from the natural human speech, and then using a one-class SVM to train a one-class classifier to distinguish natural and converted speech. The experiments conducted on the NIST SRE 2006 database showed that the LBP-based one-class classifier is able to achieve an EER of 5% in the converted speech detection task, while the corresponding two-class SVM classifier gives an EER of 0%. The one-class countermeasure reduces the FAR of the i-vector speaker verification system from 55 to 4.1% in the face of voice conversion spoofing. The LBP-based countermeasure assumes the natural texture is distorted during the conversion process; however, if multiple-frame-based speech segments are directly used in the converted speech, the original spectral texture might be preserved, in which scenario the LBP-based countermeasures might not be effective.

Different from above approaches focusing on discriminative features, a back-end spoofing detection approach was proposed in [107] for anti-spoofing at the model level. In the method, an integrated PLDA system is used to jointly operate anti-spoofing and speaker verification in i-vector space. This allows us to use the same front-end for feature extraction for both speaker verification and anti-spoofing. The experimental results reported on NIST SRE 2006 database suggest that the proposed method generalizes well for unseen voice conversion attacks. However, only

two kinds of voice conversion attacks are considered in the work.

VI. SOFTWARES AND DATABASES

Voice conversion spoofing and anti-spoofing studies involve both voice conversion and speaker verification technologies. To conduct such studies, a broad techniques are required. To this end, we point out several software packages and databases that can be used in the anti-spoofing studies for further research purpose.

For speaker verification, ALIZE [108] is one of the most popular toolkits for speaker verification. The toolkit includes speaker modeling and score normalization techniques. The speaker modeling techniques include the classic GMM–UBM modeling technique as well as the latest state-of-the-art speaker modeling techniques, such as FA, i-vector and PLDA modeling. Bob, a signal-processing and machine-learning toolbox, also provides a number of speaker modeling techniques similar to that in ALIZE [109]. The MSR Identity Toolbox is a MATLAB toolbox for speaker recognition research [110]. This toolbox implements both the classic GMM–UBM modeling and the state-of-the-art i-vector with PLDA techniques. In addition, it also provides some feature normalization and performance evaluation modules. Aside from speaker modeling, the hidden Markov model toolkit (HTK) [111] can be used as a complementary toolbox for feature extraction to above speaker modeling toolboxes.

From the side of voice conversion, there are few toolboxes. There is an implementation of the JD-GMM voice conversion in the Festvox project[1]. The voice conversion MATLAB toolbox [112] implements several voice conversion techniques, such as frequency warping and unit selection. It also provides feature extraction and speech reconstruction modules. In addition, the speech signal processing toolkit (SPTK)[2] provides a number of feature extraction and speech reconstruction techniques. This toolbox can be used with other voice conversion toolboxes as front-end for speech analysis and reconstruction.

The NIST speaker recognition evaluation (SRE) databases are the most popular corpora in the past spoofing and anti-spoofing studies. The NIST SRE databases are the benchmarking databases for TI-SV research. For the TD-SV research, the RSR2015 [60, 113] database has been used and is found to be suitable for spoofing and anti-spoofing research [114], as it simulates the real application scenario.

VII. CONCLUSION

In this paper, we present an overview of voice conversion spoofing and anti-spoofing for speaker verification. Due to the rapid development of speaker verification technology, speaker verification systems have been deployed into real applications, such as smartphone [1]. At the same time, voice conversion technology also progresses tremendously. Therefore, the countermeasures for voice conversion spoofing attacks become an important part of speaker verification deployment. In INTERSPEECH 2013, a special session on "Spoofing and countermeasures for ASV" was organized for the first time, which shows the increasing importance and attention of this research topic given by the academia and industry.

The current studies on anti-spoofing are very preliminary because the results are reported only on selected techniques. Comprehensive studies on the effects of interaction between different voice conversion techniques and different speaker recognition regimes are expected in the near future. The comprehensive studies between voice conversion and speaker verification need a standard database consisting of various voice conversion spoofing, which requires the two research communities to work together.

The voice conversion and anti-spoofing studies can improve one another. For example, the techniques/features developed for anti-spoofing might be used to identify the weakness of voice conversion, which could be investigated to improve voice conversion techniques. On the other hand, the improved voice conversion techniques will drive the improvement of speaker verification. This could be another direction to be explored in the future work.

In practice, an attacker might also use other techniques to implement spoofing, such as replay and speech synthesis. Hence, in the future study, development of countermeasures need to take other forms of spoofing into account. This kinds of generalized countermeasures will be useful for practical anti-spoofing.

[1] http://festvox.org/
[2] http://sp-tk.sourceforge.net/

REFERENCES

[1] Lee, K.A.; Ma, B.; Li, H.: Speaker verification makes its debut in smartphone, in *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, 2013.

[2] Faundez-Zanuy, M.; Hagmüller, M.; Kubin, G.: Speaker verification security improvement by means of speech watermarking. Speech Commun., **48** (12) (2006), 1608–1619.

[3] Lindberg, J.; Blomberg, M.: Vulnerability in speaker verification-a study of technical impostor techniques, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 1999.

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

[5] Wu, Z.: Gao, S.; Chng, E.S.; Li, H.: A study on replay attack and anti-spoofing for text-dependent speaker verification, in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2014.

[6] Lau, Y.W.; Wagner, M.; Tran, D.: Vulnerability of speaker verification to voice mimicking, in *Proc. Int. Symp. Intelligent Multimedia, Video and Speech Processing*, 2004.

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

[8] Hautamäki, R.G.; Kinnunen, T.; Hautamäki, V.; Leino, T.; Laukkanen, A.-M.: I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry, in *Proc. Interspeech*, 2013.

[9] Hunt, A.J.; Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

[10] Zen, H.; Tokuda, K.; Black, A.W.: Statistical parametric speech synthesis. Speech Commun., **51** (11) (2009), 1039–1064.

[11] Qian, Y.; Soong, F.K.; Yan, Z.-J.: A unified trajectory tiling approach to high quality speech rendering. IEEE Trans. Audio, Speech, Lang. Process., **21** (1–2) (2013), 280–290.

[12] Masuko, T.; Hitotsumatsu, T.; Tokuda, K.; Kobayashi, T.: On the security of HMM-based speaker verification systems against imposture using synthetic speech, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 1999.

[13] Masuko, T.; Tokuda, K.; Kobayashi, T.: Imposture using synthetic speech against speaker verification based on spectrum and pitch, in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000.

[14] Satoh, T.; Masuko, T.; Kobayashi, T.; Tokuda, K.: A robust speaker verification system against imposture using an HMM-based speech synthesis system, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2001.

[15] De Leon, P.L.; Pucher, M.; Yamagishi, J.; Hernaez, I.; Saratxaga, I.: Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans. Audio Speech Lang. Process., **20** (8) (2012), 2280–2290.

[16] Yamagishi, J.; Kobayashi, T.; Nakano, Y.; Ogata, K.; Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process., **17** (1) (2009), 66–83.

[17] Kinnunen, T.; Li, H.: An overview of text-independent speaker recognition: from features to supervectors. Speech Commun., **52** (1) (2010), 12–40.

[18] Wu, Z.; Evans, N.; Kinnunen, T.; Yamgishi, J.; Alegre, F.; Li, H.: Spoofing and countermeasures for speaker verification: a survey. Speech Commun., **66** (2015), 130–153.

[19] Helander, E.; Schwarz, J.; Nurminen, J.; Silen, H.; Gabbouj, M.: On the impact of alignment on voice conversion performance, in *Proc. Interspeech*, 2008.

[20] Erro, D.; Moreno, A.; Bonafonte, A.: INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Trans. Audio, Speech Lang. Process., **18** (5) (2010), 944–953.

[21] Abe, M.; Nakamura, S.; Shikano, K.; Kuwabara, H.: Voice conversion through vector quantization, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988.

[22] Kain, A.; Macon, M.W.: Spectral voice conversion for text-to-speech synthesis, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[23] Stylianou, Y.; Cappé, O.; Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process., **6** (2) (1998), 131–142.

[24] Toda, T.; Black, A.W.; Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process., **15** (8) (2007), 2222–2235.

[25] Helander, E.; Virtanen, T.; Nurminen, J.; Gabbouj, M.: Voice conversion using partial least squares regression. IEEE Trans. Audio Speech Lang. Process., **18** (5) (2010), 912–921.

[26] Zen, H.; Nankaku, Y.; Tokuda, K.: Continuous stochastic feature mapping based on trajectory HMMs. IEEE Trans. Audio Speech Lang. Process., **19** (2) (2011), 417–430.

[27] Narendranath, M.; Murthy, H.A.; Rajendran, S.; Yegnanarayana, B.: Transformation of formants for voice conversion using artificial neural networks. Speech Commun., **16** (2) (1995), 207–216.

[28] Desai, S.; Raghavendra, E.V.; Yegnanarayana, B.; Black, A.W.; Prahallad, K.: Voice conversion using artificial neural networks, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[29] Wu, Z.; Chng, E.S.; Li, H.: Conditional restricted Boltzmann machine for voice conversion, in *The First IEEE China Summit and Int. Conf. on Signal and Information Processing (ChinaSIP)*, 2013.

[30] Xie, F.-L.; Qian, Y.; Fan, Y.; Soong, F.K.; Li, H.: Sequence error (SE) minimization training of neural network for voice conversion, in *Proc. Interspeech*, 2014.

[31] Chen, L.-H.; Ling, Z.-H.; Liu, L.-J.; Dai, L.-R.: Voice conversion using deep neural networks with layer-wise generative training. IEEE Trans. Audio Speech Lang. Process., **22** (12) (2014), 1859–1872.

[32] Song, P.; Bao, Y.Q.; Zhao, L.; Zou, C.R.: Voice conversion using support vector regression. Electron. Lett., **47** (18) (2011), 1045–1046.

[33] Helander, E.; Silén, H.; Virtanen, T.; Gabbouj, M.: Voice conversion using dynamic kernel partial least squares regression. IEEE Trans. Audio Speech Lang. Process., **20** (3) (2012), 806–817.

[34] Chen, Y.; Chu, M.; Chang, E.; Liu, J.; Liu, R.: Voice conversion with smoothed GMM and MAP adaptation, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2003.

[35] Wu, Z.; Virtanen, T.; Kinnunen, T.; Chng, E.S.; Li, H.: Exemplar-based voice conversion using non-negative spectrogram deconvolution, in *8th ISCA Speech Synthesis Workshop*, 2013.

[36] Mizuno, H.; Abe, M.: Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. Speech Commun., **16** (2) (1995), 153–164.

[37] Sundermann, D.; Ney, H.: VTLN-based voice conversion, in *Proc. the 3rd IEEE Int. Symp. on Signal Processing and Information Technology*, 2003.

[38] Erro, D.; Moreno, A.; Bonafonte, A.: Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process., **18** (5) (2010), 922–931.

[39] Godoy, E.; Rosec, O.; Chonavel, T.: Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. IEEE Trans. Audio Speech Lang. Process., **20** (4) (2012), 1313–1323.

[40] Erro, D.; Navas, E.; Hernaez, I.: Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio Speech Lang. Process., **21** (3) (2013), 556–566.

[41] Mohammadi, S.H.; Kain, A.: Transmutative voice conversion, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[42] Sundermann, D.; Hoge, H.; Bonafonte, A.; Ney, H.; Black, A.; Narayanan, S.: Text-independent voice conversion based on unit selection, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[43] Dutoit, T.; Holzapfel, A.; Jottrand, M.; Moinet, A.; Perez, J.; Stylianou, Y.: Towards a voice conversion system based on frame selection, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[44] Wu, Z.; Virtanen, T.; Kinnunen, T.; Chng, E.S.; Li, H.: Exemplar-based unit selection for voice conversion utilizing temporal information, in *Proc. INTERSPEECH*, 2013.

[45] Chappel, D.T.; Hansen, J.H.L.: Speaker-specific pitch contour modeling and modification, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[46] Gillett, B.; King, S.: Transforming F_0 contours, in *Eurospeech*, Geneva, September 2003, 101–104.

[47] Inanoglu, Z.: Transforming Pitch in a Voice Conversion Framework. Master's thesis, St. Edmund's College, University of Cambridge, Cambridge, 2003.

[48] Wu, Z.-Z.; Kinnunen, T.; Chng, E.S.; Li, H.: Text-independent F_0 transformation with non-parallel data for voice conversion, in *Proc. INTERSPEECH*, 2010.

[49] Lovive, D.; Barbot, N.; Boeffard, O.: Pitch and duration transformation with non-parallel data, in *Speech Prosody 2008*, Campinas, Brazil, May 2008, 111–114.

[50] Helander, E.E.; Nurminen, J.: A novel method for prosody prediction in voice conversion, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, vol. **4**, April 2007, 509–512.

[51] Wu, C.-H.; Hsia, C.-C.; Liu, T.-H.; Wang, J.-F.: Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Trans. Audio Speech Lang. Process., **14** (4) (2006), 1109–1116.

[52] Nankaku, Y.; Nakamura, K.; Toda, T.; Tokuda, K.: Spectral conversion based on statistical models including time-sequence matching, in *Proc. 6th ISCA workshop on speech synthesis*, 2007.

[53] Lolive, D.; Barbot, N.; Boeffard, O.: Pitch and duration transformation with non-parallel data, *Proc. Speech Prosody*, 2008, 111–114.

[54] Yutani, K.; Uto, Y.; Nankaku, Y.; Toda, T.; Tokuda, K.: Simultaneous conversion of duration and spectrum based on statistical models including time-sequence matching, in *Proc. Interspeech*, 2008.

[55] Campbell, J.P. Jr: Speaker recognition: a tutorial. Proc. IEEE, **85** (9) (1997), 1437–1462.

[56] Bimbot, F. *et al.*: A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process., **2004** (2004), 430–451.

[57] Jin, Q.; Zheng, T.F.: Overview of front-end features for robust speaker recognition, in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2011.

[58] Togneri, R.; Pullella, D.: An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag., **11** (2) (2011), 23–61.

[59] Li, H.; Ma, B.; Lee, K.A.: Spoken language recognition: From fundamentals to practice. Proc. IEEE, **101** (5) (2013), 1136–1159.

[60] Larcher, A.; Lee, K.A.; Ma, B.; Li, H.: Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Commun., **60** (2014), 56–77.

[61] Kinnunen, T.: Joint acoustic-modulation frequency for speaker recognition, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[62] Kinnunen, T.; Lee, K.-A.; Li, H.: Dimension reduction of the modulation spectrogram for speaker verification, in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2008.

[63] Adami, A.G.; Mihaescu, R.; Reynolds, D.A.; Godfrey, J.J.: Modeling prosodic dynamics for speaker recognition, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[64] Kajarekar, G.S.; Stolcke, E.S.; Kajarekar, S.; Venkataraman, A.: Modeling duration patterns for speaker recognition, in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2003.

[65] Shriberg, E.; Ferrer, L.; Kajarekar, S.; Venkataraman, A.; Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. Speech Commun., **46** (3) (2005), 455–472.

[66] Dehak, N.; Dumouchel, P.; Kenny, P.: Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process., **15** (7) (2007), 2095–2103.

[67] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Process., **10** (1) (2000), 19–41.

[68] Kenny, P.: Joint factor analysis of speaker and session variability: theory and algorithms, Technical Report CRIM-06/08-14, 2006.

[69] Kenny, P.; Boulianne, G.; Ouellet, P.; Dumouchel, P.: Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio Speech Lang. Process., **15** (4) (2007), 1448–1460.

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

[71] Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process., **19** (4) (2011), 788–798.

[72] Campbell, W.M.; Sturim, D.E.; Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett., **13** (5) (2006), 308–311.

[73] Cumani, S.; Brummer, N.; Burget, L.; Laface, P.: Fast discriminative speaker verification in the i-vector space, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[74] Wan, V.; Renals, S.: Speaker verification using sequence discriminant support vector machines. IEEE Trans. Speech Audio Process., **13** (2) (2005), 203–210.

[75] Solomonoff, A.; Campbell, W.M.; Boardman, I.: Advances in channel compensation for SVM speaker recognition, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[76] Burget, L.; Matejka, P.; Schwarz, P.; Glembek, O.; Cernocky, J.: Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans. Audio Speech Lang. Process., **15** (7) (2007), 1979–1986.

[77] Hatch, A.O.; Kajarekar, S.; Stolcke, A.: Within-class covariance normalization for SVM-based speaker recognition, in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2006.

[78] Farrell, K.R.; Mammone, R.J.; Assaleh, K.T.: Speaker recognition using neural networks and conventional classifiers. IEEE Trans. Speech Audio Process., **2** (1) (1994), 194–205.

[79] Xiang, B.; Berger, T.: Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Trans. Speech Audio Process., **11** (5) (2003), 447–456.

[80] Ghahabi, O.; Hernando, J.: Deep belief networks for i-vector based speaker recognition, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[81] Lei, Y.; Scheffer, N.; Ferrer, L.; McLaren, M.: A novel scheme for speaker recognition using a phonetically-aware deep neural network, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[82] Lei, Y.; Ferrer, L.; McLaren, M.; Scheffer, N.: A deep neural network speaker verification system targeting microphone speech, in *Proc. Interspeech*, 2014.

[83] McLaren, M.; Lei, Y.; Scheffer, N.; Ferrer, L.: Application of convolutional neural networks to speaker recognition in noisy conditions, in *Proc. Interspeech*, 2014.

[84] Kenny, P.; Gupta, V.; Stafylakis, T.; Ouellet, P.; Alam, J.: Deep neural networks for extracting Baum–Welch statistics for speaker recognition, in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2014.

[85] Ghahabi, O.; Hernando, J.: i-Vector modeling with deep belief networks for multi-session speaker recognition, in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2014.

[86] Hautamaki, V.; Kinnunen, T.; Sedlak, F.; Lee, K.A.; Ma, B.; Li, H.: Sparse classifier fusion for speaker verification. IEEE Trans. Audio Speech Lang. Process., **21** (8) (2013), 1622–1631.

[87] Hasan, T.; Sadjadi, S.O.; Liu, G.; Shokouhi, N.; Boril, H.; Hansen, J.H.L.: CRSS systems for 2012 NIST speaker recognition evaluation, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[88] McLaren, M.; Scheffer, N.; Graciarena, M.; Ferrer, L.; Lei, Y.: Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[89] Kons, Z.; Aronowitz, H.: Voice transformation-based spoofing of text-dependent speaker verification systems, in *Proc. Interspeech*, 2013.

[90] Wu, Z.; Kinnunen, T.; Chng, E.S.; Li, H.; Ambikairajah, E.: A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case, in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2012.

[91] Bonastre, J.-F.; Matrouf, D.; Fredouille, C.: Artificial impostor voice transformation effects on false acceptance rates, in *Proc. Interspeech*, 2007.

[92] Alegre, F.; Vipperla, R.; Evans, N.; Fauve, B.: On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals, in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2012.

[93] Kinnunen, T.; Wu, Z.-Z.; Lee, K.A.; Sedlak, F.; Chng, E.S.; Li, H.: Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[94] Aronowitz, H.; Hoory, R.; Pelecanos, J.; Nahamoo, D.: New developments in voice biometrics for user authentication, in *Proc. Interspeech*, 2011.

[95] Pellom, B.L.; Hansen, J.H.L.: An experimental study of speaker verification sensitivity to computer voice-altered imposters, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[96] Perrot, P.; Aversano, G.; Blouet, R.; Charbit, M.; Chollet, G.: Voice forgery using ALISP: indexation in a client memory, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[97] Matrouf, D.; Bonastre, J.-F.; Fredouille, C.: Effect of speech transformation on impostor acceptance, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[98] Alegre, F.; Vipperla, R.; Evans, N.: Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals, in *Proc. Interspeech*, 2012.

[99] Erro, D.; Navas, E.; Hernaez, I.: Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio Speech Lang. Process. **21** (3) (2013), 556–566.

[100] Wu, Z.; Chng, E.S.; Li, H.: Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition, in *Proc. Interspeech*, 2012.

[101] Wu, Z.; Xiao, X.; Chng, E.S.; Li, H.: Synthetic speech detection using temporal modulation feature, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[102] Takamichi, S.; Toda, T.; Neubig, G.; Sakti, S.; Nakamura, S.: A post-filter to modify the modulation spectrum in HMM-based speech synthesis, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[103] Alegre, F.; Vipperla, R.; Amehraye, A.; Evans, N.: A new speaker verification spoofing countermeasure based on local binary patterns, in *Proc. Interspeech*, 2013.

[104] Ahonen, T.; Hadid, A.; Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell., **28** (12) (2006), 2037–2041.

[105] Chingovska, I.; Anjos, A.; Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing, in *Proc. Int. Conf. of Biometrics Special Interest Group (BIOSIG)*., IEEE, 2012, 1–7.

[106] Alegre, F.; Amehraye, A.; Evans, N.: A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns, in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[107] Khoury, E.; Kinnunen, T.; Sizov, A.; Wu, Z.; Marcel, S.: Introducing i-vectors for joint anti-spoofing and speaker verification, in *Proc. Interspeech*, 2014.

[108] Larcher, A. *et al.*: Alize 3.0-open source toolkit for state-of-the-art speaker recognition, in *Proc. Interspeech*, 2013.

[109] Anjos, A.; El-Shafey, L.; Wallace, R.; Günther, M.; McCool, C.; Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers, in *Proc. 20th ACM Int. Conf. on Multimedia*, 2012.

[110] Sadjadi, S.O.; Slaney, M.; Heck, L.: MSR identity toolbox v1.0: A matlab toolbox for speaker-recognition research, in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter*, 2013.

[111] Young, S. *et al.*: The HTK Book (for HTK Version 3.4), Engineering Department, Cambridge University, 2006.

[112] Sündermann, D.: Voice Conversion MATLAB Toolbox, Technical Report, Siemens Corporate Technology, Munich, Germany, 2007.

[113] Larcher, A.; Lee, K.-A.; Ma, B.; Li, H.: RSR2015: database for text-dependent speaker verification using multiple pass-phrases, in *Proc. Interspeech*, 2012.

[114] Wu, Z.; Larcher, A.; Lee, K.A.; Chng, E.S.; Kinnunen, T.; Li, H.: Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints, in *Proc. Interspeech*, 2013.

Zhizheng Wu is now a post-doctoral research associate in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, United Kingdom. Before joining CSTR, he was a Ph.D. Student in Nanyang Technological University, Singapore from 2010 to 2014. He was awarded the best paper award by Asia-Pacific Signal and Information Processing Association for this anti-spoofing work in Dec. 2012. His research interests include speech synthesis, voice conversion and speaker verification.

Haizhou Li received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently the Principal Scientist, Department Head of Human Language Technology in the Institute for Infocomm Research (I^2R), Singapore. He is also an Adjunct Professor at the National University of Singapore and a Conjoint Professor at the University of New South Wales, Australia. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing.

Prior to joining I^2R, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), a Research Manager at the Apple-ISS Research Centre (1996–1998), a Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and the Vice President in InfoTalk Corp. Ltd. (2001–2003).

Dr Li is currently the President of Asia Pacific Signal and Information Processing Association (2015–2016). He is also the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015–2017). He has served in the Editorial Board of Computer Speech and Language (2012–2014). He is an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the Vice President of the International Speech Communication Association (2013–2014). He was the General Chair of ACL 2012 and INTERSPEECH 2014.

Dr Li is a Fellow of the IEEE. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one the two Nokia Visiting Professors in 2009 by the Nokia Foundation.