Speaker Recognition Anti-Spoofing

Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi,
Zhizheng Wu, Federico Alegre and Phillip De Leon

Abstract Progress in the development of spoofing countermeasures for automatic speaker recognition is less advanced than equivalent work related to other biometric modalities. This chapter outlines the potential for even state-of-the-art automatic speaker recognition systems to be spoofed. While the use of a multitude of different datasets, protocols and metrics complicates the meaningful comparison of different vulnerabilities, we review previous work related to impersonation, replay, speech synthesis and voice conversion spoofing attacks. The article also presents an analysis of the early work to develop spoofing countermeasures. The literature shows that there is significant potential for automatic speaker verification systems to be spoofed, that significant further work is required to develop generalised countermeasures, that there is a need for standard datasets, evaluation protocols and metrics and that greater emphasis should be placed on text-dependent scenarios.

Nicholas Evans and Federico Alegre
EURECOM, Department of Multimedia Communications, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France, e-mail: evans@eurecom.fr,alegre@eurecom.fr
Tomi Kinnunen
University of Eastern Finland (UEF), Speech and Image Processing Unit, School of Computing, P.O. Box 111, FI-80101 Joensuu, FINLAND, e-mail: tkinnu@cs.uef.fi
Junichi Yamaghishi
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan and University of Edinburgh, 10 Crichton Street Edinburgh, EH8 9AB, United Kingdom, e-mail: jyamagis@inf.ed.ac.uk
Zhizheng Wu
Nanyang Technological University (NTU), Emerging Research Lab, School of Computer Engineering, N4-B1a-02 C2I, Nanyang Avenue, Singapore 639798, e-mail: wuzz@ntu.edu.sg
Phillip De Leon
New Mexico State University, Klipsch School of Elect. & Comp. Eng., Box 30001, Department 3-O, Las Cruces, New Mexico 88003-8001, USA, e-mail: pdeleon@nmsu.edu

1 Introduction

As one of our primary methods of communication, the speech modality has natural appeal as a biometric in one of two different scenarios: *text-independent* and *text-dependent*. While text-dependent automatic speaker verification (ASV) systems use fixed or randomly prompted utterances with known text content, text-independent recognisers operate on arbitrary utterances, possibly spoken in different languages. Text-independent methods are best suited to surveillance scenarios where speech signals are likely to originate from non-cooperative speakers. In authentication scenarios, where cooperation can be readily assumed, text-dependent ASV is generally more appropriate since better performance can then be achieved with shorter utterances. On the other hand, text-independent recognisers are also used for authentication in call-centre applications such as caller verification in telephone banking[1]. On account of its utility in surveillance applications, evaluation sponsorship and dataset availability, text-independent ASV dominates the field.

The potential for ASV to be spoofed is now well recognised [28]. Since speaker recognition is commonly used in telephony or other unattended, distributed scenarios without human supervision, speech is arguably more prone to malicious interference or manipulation than other biometric signals. However, while spoofing is relevant to authentication scenarios and therefore text-dependent ASV, almost all prior work has been performed on text-independent datasets more suited to surveillance. While this observation most likely reflects the absence of viable text-dependent datasets in the recent past, progress in the development of spoofing countermeasures for ASV is lagging behind that in other biometric modalities[2].

Nonetheless there is growing interest to assess the vulnerabilities of ASV to spoofing and new initiatives to develop countermeasures [28]. This article reviews the past work which is predominantly text-independent. While the use of different datasets, protocols and metrics hinders such a task, we aim to describe and analyse four different spoofing attacks considered thus far: impersonation, replay, speech synthesis and voice conversion. Countermeasures for all four spoofing attacks are also reviewed and we discuss the directions which must be taken in future work to address weaknesses in the current research methodology and to properly protect ASV systems from the spoofing threat.

2 Automatic speaker verification

This section describes state-of-the-art approaches to text-independent automatic speaker verification (ASV) and their potential vulnerabilities to spoofing.

---

[1] `http://www.nuance.com/landing-pages/products/voicebiometrics/freespeech.asp`

[2] `http://www.tabularasa-euproject.org/`

2.1 Feature extraction

Since speech signals are non-stationary, features are commonly extracted from short-term segments (frames) of 20-30 ms in duration. Typically, mel-frequency cepstral coefficient (MFCC), linear predictive cepstral coefficient (LPCC) or perceptual linear prediction (PLP) features are used as a descriptor of the short-term power spectrum. These are usually appended with their time derivative coefficients (deltas and double-deltas) and they undergo various normalisations such as global mean removal or short-term Gaussianization or feature warping [68]. In addition to spectral features, prosodic and high-level features have been studied extensively [81, 22, 82], achieving comparable results to state-of-the-art spectral recognisers [50]. For more details regarding popular feature representations used in ASV, readers are referred to [46].

The literature shows that ASV systems based on both spectral and prosodic features are vulnerable to spoofing. As described in Section 3, state-of-the-art voice conversion and statistical parametric speech synthesisers may also use mel-cepstral and linear prediction representations; spectral recognisers can be particularly vulnerable to synthesis and conversion attacks which use 'matched' parameterisations. Recognisers which utilise prosodic parameterisations are in turn vulnerable to human impersonation.

2.2 Modelling and classification

Approaches to ASV generally focus on modelling the long-term distribution of spectral vectors. To this end, the Gaussian mixture model (GMM) [74, 75] has become the *de facto* modelling technique. Early ASV systems used maximum likelihood (ML) [74] and maximum a posteriori (MAP) [75] training. In the latter case, a speaker-dependent GMM is obtained from the adaptation of a previously-trained universal background model (UBM). Adapted GMM mean *supervectors* obtained in this way were combined with support vector machine (SVM) classifiers in [14]. This idea lead to the development of many successful speaker model normalisation techniques including nuisance attribute projection (NAP) [83, 13] and within-class covariance normalisation (WCCN) [34]. These techniques aim to compensate for intersession variation, namely differences in supervectors corresponding to the same speaker caused by channel or session mismatch.

Parallel to the development of SVM-based discriminative models, generative factor analysis models were pioneered in [43, 44, 45]. In particular, *joint factor analysis* (JFA) [43] can improve ASV performance by incorporating distinct speaker and channel subspace models. These subspace models require the estimation of various hyper-parameters using labelled utterances. Subsequently, JFA evolved into a much-simplified model that is now the state-of-the-art. The so-called *total variability model* or 'i-vector' representation [21] uses latent variable vectors of low-dimension (typically 200 to 600) to represent an arbitrary utterance. Unlike JFA, the

training of an i-vector extractor is essentially an unsupervised process which leads to only one subspace model. Accordingly it can be viewed as a approach to dimensionality reduction, while compensation for session, environment and other nuisance factors are applied in the computationally light back-end classification. To this end, *probabilistic linear discriminant analysis* (PLDA) [54] with length-normalised i-vectors [32] has proven particularly effective.

Being based on the transformation of short-term cepstra, conversion and synthesis techniques also induce a form of 'channel shift'. Since they aim to attenuate channel effects, approaches to intersession compensation may present vulnerabilities to spoofing through the potential to confuse spoofed speech with channel-shifted speech of a target speaker. However, even if there is some evidence to the contrary, i.e. that recognisers employing intersession compensation might be intrinsically more robust to voice conversion attacks [47], all have their roots in the standard GMM and independent spectral observations. Neither utilises time sequence information, a key characteristic of speech which might otherwise afford some protection from spoofing.

2.3 System fusion

In addition to the development of increasingly robust models and classifiers, there is a significant emphasis within the ASV community on the study of *classifier fusion*. This is based on the assumption that independently trained recognisers capture different aspects of the speech signal not covered by any individual classifier. Fusion also provides a convenient vehicle for large-scale research collaborations promoting independent classifier development and benchmarking [76]. Different classifiers can involve different features, classifiers, or hyper-parameter training sets [12]. A simple, yet robust approach to fusion involves the weighted summation of the base classifier scores, where the weights are optimised according to a logistic regression cost function. For recent trends in fusion, readers are referred to [35].

While we are unaware of any spoofing or anti-spoofing studies on fused ASV systems, some insight into their likely utility can be gained from related work in fused, multi-modal biometric systems; whether the scores originate from different biometric modalities or sub-classifiers applied to the same biometric trait makes little difference. A common claim is that multi-biometric systems should be inherently resistant to spoofing since an impostor is less likely to succeed in spoofing *all* the different subsystems. We note, however, that [2] suggests it might suffice to spoof only *one* modality under a score fusion setting in the case where the spoofing of a single, significantly weighted sub-system is particularly effective.

3 Spoofing and countermeasures

Spoofing attacks are performed on a biometric system at the sensor or acquisition level to bias score distributions toward those of genuine clients, thus provoking increases in the false acceptance rate (FAR). This section reviews past work to evaluate vulnerabilities and to develop spoofing countermeasures. We consider impersonation, replay, speech synthesis and voice conversion.

3.1 Impersonation

Impersonation refers to spoofing attacks whereby a speaker attempts to imitate the speech of another speaker and is one of the most obvious forms of spoofing and earliest studied.

3.1.1 Spoofing

The work in [52] showed that impersonators can readily adapt their voice to overcome ASV, but only when their natural voice is already similar to that of the target (the *closest* targets were selected from YOHO corpus using an ASV system). Further work in [51] showed that impersonation increased FAR rates from close to 0% to between 10% and 60%. Linguistic expertise was not found to be useful, except in cases when the voice of the target speaker was very different to that of the impersonator. However, contradictory findings reported in [58] suggest that even while professional imitators are better impersonators than average people, they are *unable* to spoof an ASV system.

In addition to spoofing studies, impersonation has been a subject in acoustic-phonetic studies [25, 107, 29]. These have shown that imitators tend to be effective in mimicking long-term prosodic patterns and the speaking rate, though it is less clear that they are as effective in mimicking formant and other spectral characteristics. For instance, the imitator involved in the studies reported in [25] was not successful in translating his formant frequencies towards the target, whereas the opposite is reported in [48].

Characteristic to all studies involving impersonation is the use of relatively few speakers, different languages and ASV systems. The target speakers involved in such studies are also often public figures or celebrities and it is difficult to collect technically comparable material from both the impersonator and the target. These aspects of the past work makes it difficult to conclude whether or not impersonation poses a genuine threat. Since impersonation is thought to involve mostly the mimicking of prosodic and stylistic cues, it is perhaps considered more effective in fooling human listeners than today's state-of-the-art ASV systems [70].

3.1.2 Countermeasures

While the threat of impersonation is not fully understood due to limited studies involving small datasets, it is perhaps not surprising that there is no prior work to investigate countermeasures against impersonation. If the threat is proven to be genuine, then the design of appropriate countermeasures might be challenging. Unlike the spoofing attacks discussed below, all of which can be assumed to leave traces of the physical properties of the recording and playback devices, or signal processing artifacts from synthesis or conversion systems, impersonators are live human beings who produce entirely natural speech.

3.2 Replay

Replay attacks involve the presentation of previously-recorded speech from a genuine client in the form of continuous speech recordings, or samples resulting from the concatenation of shorter segments. Replay is a relatively low-technology attack within the grasp of any potential attacker even without specialised knowledge in speech processing. The availability of inexpensive, high quality recording devices and digital audio editing software might suggest that replay is both effective and difficult to detect.

3.2.1 Spoofing

In contrast to research involving speech synthesis and voice conversion spoofing attacks where large datasets are generally used for assessment, e.g. NIST datasets, all the past work to assess vulnerabilities to replay attacks relates to small, often purpose-collected datasets, typically involving no more than 15 speakers. While results generated with such small datasets have low statistical significance, differences between baseline performance and that under spoofing highlight the vulnerability.

The vulnerability of ASV systems to replay attacks was first investigated in a text-dependent scenario [55] where the concatenation of recorded digits were tested against a hidden Markov model (HMM) based ASV system. Results showed an increase in the FAR (EER threshold) from 1% to 89% for male speakers and from 5% to 100% for female speakers.

The work in [90] investigated text-independent ASV vulnerabilities through the replaying of far-field recorded speech in a mobile telephony scenario where signals were transmitted by analogue and digital telephone channels. Using a baseline ASV system based on JFA, their work showed an increase in the EER of 1% to almost 70% when impostor accesses were replaced by replayed spoof attacks. A physical access scenario was considered in [92]. While the baseline performance of their GMM-UBM ASV system was not reported, experiments showed that replay attacks produced an FAR of 93%.

3.2.2 Countermeasures

A countermeasure for replay attack detection in the case of text-dependent ASV was reported in [80]. The approach is based upon the comparison of new access samples with stored instances of past accesses. New accesses which are deemed too similar to previous access attempts are identified as replay attacks. A large number of different experiments, all relating to a telephony scenario, showed that the countermeasures succeeded in lowering the EER in most of the experiments performed.

While some form of text-dependent or challenge-response countermeasure is usually used to prevent replay-attacks, text-independent solutions have also been investigated. The same authors in [90] showed that it is possible to detect replay attacks by measuring the channel differences caused by far-field recording [91]. While they show spoof detection error rates of less than 10% it is feasible that today's state-of-the-art approaches to channel compensation will render some ASV systems still vulnerable.

Two different replay attack countermeasures are compared in [92]. Both are based on the detection of differences in channel characteristics expected between licit and spoofed access attempts. Replay attacks incur channel noise from both the recording device and the loudspeaker used for replay and thus the detection of channel effects beyond those introduced by the recording device of the ASV system thus serves as an indicator of replay. The performance of a baseline GMM-UBM system with an EER 40% under spoofing attack falls to 29% with the first countermeasure and a more respectable EER of 10% with the second countermeasure.

3.3 Speech synthesis

Speech synthesis, commonly referred to as text-to-speech (TTS), is a technique for generating intelligible, natural-sounding artificial speech for any arbitrary text. Speech synthesis is used widely in various applications including in-car navigation systems, e-book readers, voice-over functions for the visually impaired, and communication aids for the speech impaired. More recent applications include spoken dialogue systems, communicative robots, singing speech synthesisers, and speech-to-speech translation systems.

Typical speech synthesis systems have two main components: text analysis and speech waveform generation, which are sometimes referred to as the *front-end* and *back-end*, respectively. In the text analysis component, input text is converted into a linguistic specification consisting of elements such as phonemes. In the speech waveform generation component, speech waveforms are generated from the produced linguistic specification.

There are four major approaches to speech waveform generation. In the early 1970s, the speech waveform generation component used very low dimensional acoustic parameters for each phoneme, such as formants, corresponding to vocal tract resonances with hand-crafted acoustic rules [49]. In the 1980s, the speech

waveform generation component used a small database of phoneme units called ‘diphones’ (the second half of one phone plus the first half of the following phone) and concatenated them according to the given phoneme sequence by applying signal processing, such as linear predictive (LP) analysis, to the units [65]. In the 1990s, larger speech databases were collected and used to select more appropriate speech units that match both phonemes and other linguistic contexts such as lexical stress and pitch accent in order to generate high-quality natural sounding synthetic speech with appropriate prosody. This approach is generally referred to as ‘unit selection,’ and is used in many speech synthesis systems, including commercial products [40, 11, 24, 9, 17]. In the late 1990s another data-driven approach emerged, ‘Statistical parametric speech synthesis,’ and has grown in popularity in recent years [103, 56, 10, 105]. In this approach, several acoustic parameters are modelled using a time-series stochastic generative model, typically a hidden Markov model (HMM). HMMs represent not only the phoneme sequences but also various contexts of the linguistic specification in a similar way to the unit selection approach. Acoustic parameters generated from HMMs and selected according to the linguistic specification are used to drive a vocoder (a simplified speech production model with which speech is represented by vocal tract and excitation parameters) in order to generate a speech waveform.

The first three approaches are unlikely to be effective in ASV spoofing since they do not provide for the synthesis of speaker-specific formant characteristics. Furthermore, diphone or unit selection approaches generally require a speaker-specific database that covers all the diphones or relatively large amounts of speaker-specific data with carefully prepared transcripts. In contrast, state-of-the-art HMM-based speech synthesisers [106, 102] can learn individualised speech models from relatively little speaker-specific data by adapting background models derived from other speakers based on the standard model adaptation techniques drawn from speech recognition, i.e. maximum likelihood linear regression (MLLR) [53, 93].

3.3.1 Spoofing

There is a considerable volume of research in the literature which has demonstrated the vulnerability of ASV to synthetic voices generated with a variety of approaches to speech synthesis. Experiments using formant, diphone, and unit-selection based synthetic speech in addition to the simple cut-and-paste of speech waveforms have been reported [55, 30, 90].

ASV vulnerabilities to HMM-based synthetic speech were first demonstrated over a decade ago [60] using an HMM-based, text-prompted ASV system [64] and an HMM-based synthesiser where acoustic models were adapted to specific human speakers [61, 62]. The ASV system scored feature vectors against speaker and background models composed of concatenated phoneme models. When tested with human speech the ASV system achieved an FAR of 0% and an FRR of 7%. When subjected to spoofing attacks with synthetic speech, the FAR increased to over 70%, however this work involved only 20 speakers.

Large-scale experiments using the Wall Street Journal corpus containing 284 speakers and two different ASV systems (GMM-UBM and SVM using Gaussian supervectors) was reported in [19]. Using a state-of-the-art HMM-based speech synthesiser, the FAR was shown to rise to 86% and 81% for the GMM-UBM and SVM systems, respectively. Spoofing experiments using HMM-based synthetic speech against a forensics speaker verification tool *BATVOX* was also reported in [31] with similar findings. Today's state-of-the-art speech synthesisers thus present a genuine threat to ASV.

3.3.2 Countermeasures

Only a small number of attempts to discriminate synthetic speech from natural speech have been investigated and there is currently no general solution which is independent from specific speech synthesis methods. Previous work has demonstrated the successful detection of synthetic speech based on prior knowledge of the acoustic differences of specific speech synthesisers, such as the dynamic ranges of spectral parameters at the utterance level [79] and variance of higher order parts of mel-cepstral coefficients [15].

There are some attempts which focus on acoustic differences between vocoders and natural speech. Since the human auditory system is known to be relatively insensitive to phase [73], vocoders are typically based on a minimum-phase vocal tract model. This simplification leads to differences in the phase spectra between human and synthetic speech, differences which can be utilised for discrimination [19, 95].

Based on the difficulty in reliable prosody modelling in both unit selection and statistical parametric speech synthesis, other approaches to synthetic speech detection use F0 statistics [67, 20]. F0 patterns generated for the statistical parametric speech synthesis approach tend to be over-smoothed and the unit selection approach frequently exhibits 'F0 jumps' at concatenation points of speech units.

3.4 Voice conversion

Voice conversion is a sub-domain of voice transformation [85] which aims to convert one speaker's voice towards that of another. The field has attracted increasing interest in the context of ASV vulnerabilities for over a decade [69]. Unlike TTS, which requires text input, voice conversion operates directly on speech samples. In particular, the goal is to transform according to a conversion function $\mathscr{F}$ the feature vectors ($\mathbf{x}$) corresponding to speech from a source speaker (spoofer) to that they are closer to those of target a speaker ($\mathbf{y}$):

$$\mathbf{y} = \mathscr{F}(\mathbf{x}, \theta). \tag{1}$$

Most voice conversion approaches adopt a training phase which requires frame-aligned pairs $\{(\mathbf{x}_t, \mathbf{y}_t)\}$ in order to learn the transformation parameters θ. Frame alignment is usually achieved using dynamic time warping (DTW) on *parallel* source-target training utterances with identical text content. The trained conversion function is then applied to new source utterances of arbitrary text content at runtime.

A large number of specific conversion approaches have been reported. One of the earliest and simplest techniques employs vector quantisation (VQ) with codebooks [1] or segmental codebooks [8] of paired source-target frame vectors to represent the conversion function. However, VQ introduces frame-to-frame discontinuity problems. Among the more recent conversion methods, *joint density Gaussian mixture model* (JD-GMM) [42, 86, 89] has become a standard baseline method. It achieves smooth feature transformations using a local linear transformation. Despite its popularity, known problems of JD-GMM include over-smoothing [72, 16, 41] and over-fitting [38, 71] which has led to the development of alternative linear conversion methods such as partial least square (PLS) regression [38], tensor representation [78], a trajectory hidden Markov model [104], a mixture of factor analysers [97], local linear transformation [72] and a noisy channel model [77]. Non-linear approaches, including artificial neural networks [66, 23], support vector regression [84], kernel partial least square [37], and conditional restricted Boltzmann machines [96], have also been studied. As alternatives to data-driven conversion, frequency warping techniques [88, 26, 27] have also attracted attention.

The approaches to voice conversion considered above are usually applied to the transformation of spectral envelope features, though the conversion of prosodic features such as fundamental frequency [33, 94, 39, 101] and duration [94, 57] has also been studied. In contrast to parametric methods, unit selection approaches can be applied directly to feature vectors coming from the target speaker to synthesise converted speech [87]. Since they use target speaker data directly, unit-selection approaches arguably pose a greater risk to ASV than statistical approaches [99].

In general, only the most straightforward of the spectral conversion methods have been utilised in ASV vulnerability studies. Even when trained using a non-parallel technique and non-ideal telephony data, the baseline JD-GMM approach, which produces over-smooth speech with audible artifacts, is shown to increase significantly the FAR of modern ASV systems [47, 99]; unlike the human ear, current recognisers are essentially ‘deaf’ to obvious conversion artifacts caused by imperfect signal analysis-synthesis models and poorly trained conversion functions.

3.4.1 Spoofing

When applied to spoofing, voice conversion aims to synthesise a new speech signal such that features extracted for ASV are close in some sense to the target speaker. Some of the first work relevant to text-independent ASV spoofing includes that in [70, 63]. The work in [70] showed that a baseline EER increased from 16% to 26% as a result of voice conversion which also converted prosodic aspects not mod-

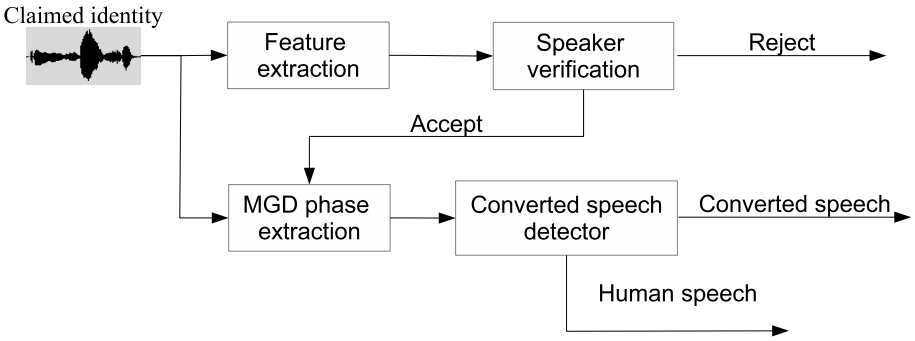

Fig. 1 An example of a spoofed speech detector combined with speaker verification [98]. Based on prior knowledge that many analysis-synthesis modules used in voice conversion and TTS systems discard natural speech phase, phase characteristics parametrised via the modified group delay function (MGDF) can be used for discriminating natural and synthetic speech.

elled in typical ASV systems. The work in [63] investigated the probabilistic mapping of a speaker's vocal tract information towards that of another, target speaker using a pair of tied speaker models, one of ASV features and another of filtering coefficients. This work targeted the conversion of spectral-slope parameters. The work showed that a baseline EER of 10% increased to over 60% when all impostor test samples were replaced with converted voice. In addition, signals subjected to voice conversion did not exhibit any perceivable artifacts indicative of manipulation.

The work in [47] investigated ASV vulnerabilities using a popular approach to voice conversion [42] based on JD-GMMs, which requires a parallel training corpus for both source and target speakers. Even if converted speech would be easily detectable by human listeners, experiments involving five different ASV systems showed universal susceptibility to spoofing. The FAR of the most robust, JFA system increased from 3% to over 17%.

Other work relevant to voice conversion includes attacks referred to as artificial signals. It was noted in [6] that certain short intervals of converted speech yield extremely high scores or likelihoods. Such intervals are not representative of intelligible speech but they are nonetheless effective in overcoming typical text-independent ASV systems which lack any form of speech quality assessment. The work in [6] showed that artificial signals optimised with a genetic algorithm provoke increases in the EER from 10% to almost 80% for a GMM-UBM system and from 5% to almost 65% for a factor analysis (FA) system.

3.4.2 Countermeasures

Some of the first work to detect converted voice draws on related work in synthetic speech detection [18]. While the proposed cosine phase and modified group delay function (MGDF) countermeasures proposed in [95, 98] are effective in detecting

Table 1 A summary of the four approaches to ASV spoofing, their expected accessibility and risk.

Spoofing technique	Description	Accessibility (practicality)	Effectiveness (risk) Text-indep.	Text-dep.
Impersonation [52, 58, 70, 36]	Human voice mimic	Low	Low/unknown	Low/unknown
Replay [55, 90]	Replay of pre-recorded utterance	High	High	Low (rand. phrase) to high (fixed phrase)
Text-to-speech [60, 64, 19]	Speaker-specific speech generation from text input	Medium (now) to high (future)	High	High
Voice conversion [70, 63, 47, 6]	Speaker identity conversion using speech only	Medium (now) to high (future)	High	High

spoofed speech (see Fig. 1), they are unlikely to detect converted voice with real-speech phase [63].

Two approaches to artificial signal detection are reported in [7]. Experimental work shows that supervector-based SVM classifiers are naturally robust to such attacks whereas all spoofing attacks can be detected using an utterance-level variability feature which detects the absence of natural, dynamic variability characteristic of genuine speech. An alternative approach based on voice quality analysis is less dependent on explicit knowledge of the attack but less effective in detecting attacks.

A related approach to detect converted voice is proposed in [4]. Probabilistic mappings between source and target speaker models are shown to yield converted speech with less short-term variability than genuine speech. The thresholded, average pair-wise distance between consecutive feature vectors is used to detect converted voice with an EER of under 3%.

Due to fact that current analysis-synthesis techniques operate at the short-term frame level, the use of temporal magnitude/phase modulation features, a form of long-term feature, are proposed in [100] to detect both speech synthesis and voice conversion spoofing attacks. Another form of long-term feature is reported in [5]. The approach is based on the local binary pattern (LBP) analysis of sequences of acoustic vectors and is successful in detecting converted voice. Interestingly, the approach is less reliant on prior knowledge and can also detect different spoofing attacks, examples of which were not used for training or optimisation.

3.5 Summary

As shown above, ASV spoofing and countermeasures have been studied with a multitude of different datasets, evaluation protocols and metrics, with highly diverse experimental designs, different ASV recognisers and with different approaches to

spoofing; the lack of any commonality makes the comparison of results, vulnerabilities and countermeasure performance an extremely challenging task. Drawing carefully upon the literature and the authors' own experience with various spoofing approaches, we have nevertheless made such an attempt. Table 1 aims to summarise the threat of spoofing for the four approaches considered above. *Accessibility (practicality)* reflects whether the threat is available to the masses or limited to the technically-knowledgeable. *Effectiveness (risk)*, in turn, reflects the success of each approach in provoking higher false acceptance rates.

Although some studies have shown that impersonation can fool ASV recognisers, in practice, the effectiveness seems to depend both on the skill of the impersonator, the similarity of the attacker's voice to that of the target speaker, and on the recogniser itself. Replay attacks are highly effective in the case of text-independent ASV and fixed-phrase text-independent systems. Even if the effectiveness is reduced in the case of randomised, phrase-prompted text-dependent systems, replay attacks are the most accessible approach to spoofing, requiring only a recording and playback device such as a tape recorder or a smart phone.

Speech synthesis and voice conversion attacks pose the greatest risk. While voice conversion systems are not yet commercially available, both free and commercial text-to-speech (TTS) systems with pre-trained voice profiles are widely available, even if commercial off-the-shelf (COTS) systems do not include the functionality for adaptation to specific target voices. While accessibility is therefore medium in the short term, speaker adaptation remains a highly active research topic. It is thus only a matter of time until flexible, speaker-adapted synthesis and conversion systems become readily available. Then, both effectiveness and accessibility should be considered high.

4 Discussion

In this section we discuss current approaches to evaluation and some weaknesses in the current evaluation methodology. While much of the following is not necessarily specific to the speech modality, with research in spoofing and countermeasures in ASV lagging behind that related to other biometric modalities, the discussion below is particularly pertinent.

4.1 Protocols and metrics

While countermeasures can be integrated into existing ASV systems, they are most often implemented as independent modules which allow for the *explicit detection* of spoofing attacks. The most common approach in this case is to concatenate the two classifiers in series.

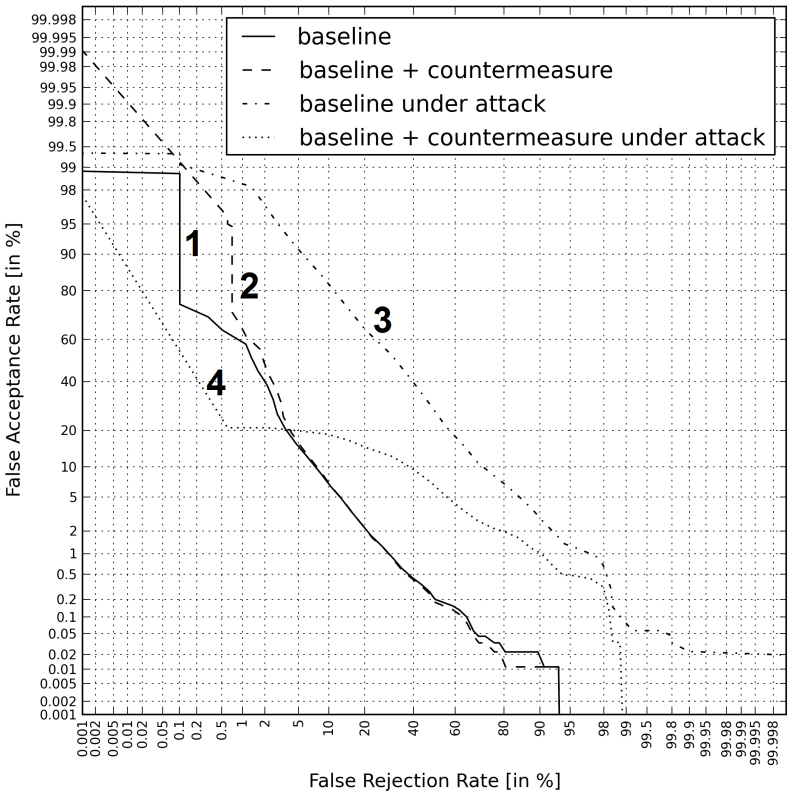

Fig. 2 An example of four DET profiles needed to analyse vulnerabilities to spoofing and countermeasure performance, both on licit and spoofed access attempts. Results correspond to spoofing attacks using synthetic speech and a standard GMM-UBM classifier assessed on the male subset of the NIST'06 SRE dataset.

The assessment of countermeasure performance on its own is relatively straightforward; results are readily analysed with standard detection error trade-off (DET) profiles [59] and related metrics. It is often of interest, however, that the assessment reflects their impact on ASV performance. Assessment is then non-trivial and calls for the joint optimisation of combined classifiers. Results furthermore reflect the performance of specific ASV systems. As described in Section 3, there are currently no standard evaluation protocols, metrics or ASV systems which might otherwise be used to conduct evaluations. There is a thus a need to define such standards in the future.

Candidate standards are being drafted within the scope of the EU FP7 TABULA RASA project[3]. Here, independent countermeasures preceding biometric verification are optimised at three different operating points where thresholds are set to obtain FARs (the probability of labelling a genuine access as a spoofing attack) of 1%, 5% or 10%. Samples labelled as genuine accesses are then passed to the verification system[4]. Performance is assessed using four different DET profiles[5], examples of which are illustrated in Figure 2. The four profiles illustrate performance of the baseline system with zero-effort impostors, the baseline system with active countermeasures, the baseline system where all impostor accesses are replaced with spoofing attacks and, finally, the baseline system with spoofing attacks and active countermeasures.

Consideration of all four profiles is needed to gauge the impact of countermeasure performance on licit transactions (any deterioration in false rejection – difference between 1st and 2nd profiles) and improved robustness to spoofing (improvements in false acceptance – difference between 3rd and 4th profiles). While the interpretation of such profiles is trivial, different plots are obtained for each countermeasure operating point. Further work is required to design intuitive, universal metrics which represent the performance of spoofing countermeasures when combined with ASV.

4.2 Datasets

While some work has shown the potential for detecting spoofing without prior knowledge or training data indicative of a specific attack [95, 5, 3], all previous work is based on some implicit prior knowledge, i.e. the nature of the spoofing attack and/or the targeted ASV system is known. While training and evaluation data with known spoofing attacks might be useful to develop and optimise appropriate countermeasures, the precise nature of spoofing attacks can never be known in practice. Estimates of countermeasure performance so obtained should thus be considered at best optimistic. Furthermore, the majority of the past work was also conducted under matched conditions, i.e. data used to learn target models and that used to effect spoofing were collected in the same or similar acoustic environment and over the same or similar channel. The performance of spoofing countermeasures when subjected to realistic session variability is then unknown.

While much of the past work already uses standard datasets, e.g. NIST SRE data, spoofed samples are obtained by treating them with non-standard algorithms. Standard datasets containing both licit transactions and spoofed speech from a multitude

[3] `http://www.tabularasa-euproject.org/`

[4] In practice samples labelled as spoofing attacks cannot be fully discarded since so doing would unduly influence false reject and false acceptance rates calculated as a percentage of all accesses.

[5] Produced with the TABULA RASA Scoretoolkit: `http://publications.idiap.ch/downloads/reports/2012/Anjos_Idiap-Com-02-2012.pdf`

of different spoofing algorithms and with realistic session variability are therefore needed to reduce the use of prior knowledge, to improve the comparability of different countermeasures and their performance against varied spoofing attacks. Collaboration with colleagues in other speech and language processing communities, e.g. voice conversion and speech synthesis, will help to assess vulnerabilities to state-of-the art spoofing attacks and also to assess countermeasures when details of the spoofing attacks are unknown. The detection of spoofing will then be considerably more challenging but more reflective of practical use cases.

5 Conclusions

This contribution reviews previous work to assess the threat from spoofing to automatic speaker verification (ASV). While there are currently no standard datasets, evaluation protocols or metrics, the study of impersonation, replay, speech synthesis and voice conversion spoofing attacks reported in this article indicate genuine vulnerabilities. We nonetheless argue that significant additional research is required before the issue of spoofing in ASV is properly understood and conclusions can be drawn.

In particular, while the situation is slowly changing, the majority of past work involves text-independent ASV, most relevant to surveillance. The spoofing threat is pertinent in authentication scenarios where text-dependent ASV might be preferred. Greater effort is therefore needed to investigate spoofing in text-dependent scenarios with particularly careful consideration being given to design appropriate datasets and protocols.

Secondly, almost all ASV spoofing countermeasures proposed thus far are dependent on training examples indicative of a specific attack. Given that the nature of spoofing attacks can never be known in practice, and with the variety in spoofing attacks being particularly high in ASV, future work should investigate new countermeasures which generalise well to unforeseen attacks. Formal evaluations with standard datasets, evaluation protocols, metrics and even standard ASV systems are also needed to address weaknesses in the current evaluation methodology.

Finally, some of the vulnerabilities discussed in this paper involve relatively high-cost, high-technology attacks. While the trend of open source software may cause this to change, such attacks are beyond the competence of the unskilled and in such case the level of vulnerability is arguably overestimated. While we have touched on this issue in this article, a more comprehensive risk-based assessment is needed to ensure such evaluations are not overly-alarmist. Indeed, the work discussed above shows that countermeasures, some of them relatively trivial, have the potential to detect spoofing attacks with manageable impacts on system usability.

Acknowledgements This work was partially supported by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU) (grant agreement number

257289), by the Academy of Finland (project no. 253120) and by EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

Index

Glossary

impersonation: a spoofing attack against automatic speaker verification whereby a speaker attempts to imitate the speech of another speaker, 5

replay: a spoofing attack against automatic speaker verification with the replaying of pre-recorded utterances of the target speaker, 6

speech synthesis: a spoofing attack against automatic speaker verification using automatically synthesised speech signals generated from arbitrary text, 7

voice conversion: a spoofing attack against automatic speaker verification using an attackers natural voice which is converted towards that of the target, 9

References

1. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 655–658. IEEE (1988)
2. Akhtar, Z., Fumera, G., Marcialis, G.L., Roli, F.: Evaluation of serial and parallel multibiometric systems under spoong attacks. In: Proc. 5th Int. Conf. on Biometrics (ICB 2012), pp. 283–288. New Delhi, India (2012)
3. 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). Washington DC, USA (2013)
4. Alegre, F., Amehraye, A., Evans, N.: Spoofing countermeasures to protect automatic speaker verification from voice conversion. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP) (2013)
5. Alegre, F., Vipperla, R., Amehraye, A., Evans, N.: A new speaker verification spoofing countermeasure based on local binary patterns. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. Lyon, France (2013)
6. Alegre, F., Vipperla, R., Evans, N., Fauve, B.: On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In: Proc. EURASIP Eur. Signal Process. Conf. (EUSIPCO). EURASIP (2012)
7. Alegre, F., Vipperla, R., Evans, N., et al.: Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. (2012)
8. Arslan, L.M.: Speaker transformation algorithm using segmental codebooks (stasc). Speech Communication **28**(3), 211–226 (1999)
9. Beutnagel, B., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A.: The AT&T Next-Gen TTS system. In: Proc. Joint ASA, EAA and DAEA Meeting, pp. 15–19 (1999)
10. Black, A.W.: CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 1762–1765 (2006)
11. Breen, A., Jackson, P.: A phonologically motivated method of selecting nonuniform units. In: Proc. IEEE Int. Conf. on Spoken Language Process. (ICSLP), pp. 2735–2738 (1998)
12. Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., Leeuwen, D., Matějka, P., Schwartz, P., Strasheim, A.: Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans. Audio, Speech and Lang. Process. **15**(7), 2072–2084 (2007)

13. Burget, L., Matějka, P., Schwarz, P., Glembek, O., Černocký, J.: Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Transactions on Audio, Speech and Language Processing **15**(7), 1979–1986 (2007)
14. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters **13**(5), 308–311 (2006)
15. Chen, L.W., Guo, W., Dai, L.R.: Speaker verification against synthetic speech. In: Proc. 7th Int. Symp. on Chinese Spoken Language Processing (ISCSLP), 2010, pp. 309–312 (29 2010-Dec. 3). DOI 10.1109/ISCSLP.2010.5684887
16. Chen, Y., Chu, M., Chang, E., Liu, J., Liu, R.: Voice conversion with smoothed GMM and MAP adaptation. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech., pp. 2413–2416 (2003)
17. Coorman, G., Fackrell, J., Rutten, P., Coile, B.: Segment selection in the L & H realspeak laboratory TTS system. In: Proc. Int. Conf. on Speech and Language Processing, pp. 395–398 (2000)
18. De Leon, P.L., Hernaez, I., Saratxaga, I., Pucher, M., Yamagishi, J.: Detection of synthetic speech for the problem of imposture. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 4844–4847. Dallas, USA (2011)
19. 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. on Audio, Speech, and Language Processing **20**(8), 2280–2290 (2012). DOI 10.1109/TASL.2012.2201472
20. De Leon, P.L., Stewart, B., Yamagishi, J.: Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. Portland, Oregon, USA (2012)
21. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech and Language Processing **19**(4), 788–798 (2011)
22. Dehak, N., Kenny, P., Dumouchel, P.: Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio, Speech and Language Processing **15**(7), 2095–2103 (2007)
23. Desai, S., Raghavendra, E.V., Yegnanarayana, B., Black, A.W., Prahallad, K.: Voice conversion using artificial neural networks. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 3893–3896. IEEE (2009)
24. Donovan, R.E., Eide, E.M.: The IBM trainable speech synthesis system. In: Proc. IEEE Int. Conf. on Spoken Language Process. (ICSLP), pp. 1703–1706 (1998)
25. Eriksson, A., Wretling, P.: How flexible is the human voice? - a case study of mimicry. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech., pp. 1043–1046 (1997). URL `http://www.ling.gu.se/~anders/papers/a1008.pdf`
26. Erro, D., Moreno, A., Bonafonte, A.: Voice conversion based on weighted frequency warping. IEEE Trans. on Audio, Speech, and Language Processing **18**(5), 922–931 (2010)
27. Erro, D., Navas, E., Hernaez, I.: Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. on Audio, Speech, and Language Processing **21**(3), 556–566 (2013)
28. Evans, N., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. Lyon, France (2013)
29. Farrús, M., Wagner, M., Anguita, J., Hernando, J.: How vulnerable are prosodic features to professional imitators? In: The Speaker and Language Recognition Workshop (Odyssey 2008). Stellenbosch, South Africa (2008)
30. Foomany, F., Hirschfield, A., Ingleby, M.: Toward a dynamic framework for security evaluation of voice verification systems. In: IEEE Toronto Int. Conf. on Science and Technology for Humanity (TIC-STH), pp. 22–27 (2009). DOI 10.1109/TIC-STH.2009.5444499
31. Galou, G.: Synthetic voice forgery in the forensic context: a short tutorial. In: Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG), pp. 1–3 (2011)

32. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of i-vector length normalization in speaker recognition systems. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 249–252. Florence, Italy (2011)
33. Gillet, B., King, S.: Transforming F0 contours. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech., pp. 101–104 (2003)
34. Hatch, A.O., Kajarekar, S., Stolcke, A.: Within-class covariance normalization for svm-based speaker recognition. In: Proc. IEEE Int. Conf. on Spoken Language Process. (ICSLP), pp. 1471–1474 (2006)
35. Hautamäki, V., Kinnunen, T., Sedlák, F., Lee, K.A., Ma, B., Li, H.: Sparse classifier fusion for speaker verification. IEEE Trans. Audio, Speech and Language Processing **21**(8), 1622–1631 (2013)
36. Hautamki, R.G., Kinnunen, T., Hautamki, V., Leino, T., Laukkanen, A.M.: I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. (2013)
37. Helander, E., Silén, H., Virtanen, T., Gabbouj, M.: Voice conversion using dynamic kernel partial least squares regression. Audio, Speech, and Language Processing, IEEE Trans. on **20**(3), 806–817 (2012)
38. Helander, E., Virtanen, T., Nurminen, J., Gabbouj, M.: Voice conversion using partial least squares regression. Audio, Speech, and Language Processing, IEEE Trans. on **18**(5), 912–921 (2010)
39. Helander, E.E., Nurminen, J.: A novel method for prosody prediction in voice conversion. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. IV–509. IEEE (2007)
40. Hunt, A., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 373–376 (1996)
41. Hwang, H.T., Tsao, Y., Wang, H.M., Wang, Y.R., Chen, S.H.: A study of mutual information for GMM-based spectral conversion. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. (2012)
42. Kain, A., Macon, M.W.: Spectral voice conversion for text-to-speech synthesis. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), vol. 1, pp. 285–288. IEEE (1998)
43. Kenny, P.: Joint factor analysis of speaker and session variability: theory and algorithms. technical report CRIM-06/08-14 (2006)
44. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech and Language Processing **15**(4), 1448–1460 (2007)
45. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of inter-speaker variability in speaker verification. IEEE Transactions on Audio, Speech and Language Processing **16**(5), 980–988 (2008)
46. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: from features to supervectors. Speech Communication **52**(1), 12–40 (2010)
47. 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. Acoust., Speech and Signal Process.(ICASSP), pp. 4401–4404. IEEE (2012)
48. Kitamura, T.: Acoustic analysis of imitated voice produced by a professional impersonator. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 813–816. Brisbane, Australia (2008)
49. Klatt, D.H.: Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America **67**, 971–995 (1980)
50. Kockmann, M., Ferrer, L., Burget, L., Černocký, J.: i-vector fusion of prosodic and cepstral features for speaker verification. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 265–268. Florence, Italy (2011)

51. Lau, Y., Tran, D., Wagner, M.: Testing voice mimicry with the yoho speaker verification corpus. In: Knowledge-Based Intelligent Information and Engineering Systems, pp. 907–907. Springer (2005)
52. Lau, Y.W., Wagner, M., Tran, D.: Vulnerability of speaker verification to voice mimicking. In: Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on, pp. 145–148. IEEE (2004)
53. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. **9**, 171–185 (1995)
54. Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J.: Probabilistic models for inference about identity. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 144–157 (2012)
55. Lindberg, J., Blomberg, M., et al.: Vulnerability in speaker verification-a study of technical impostor techniques. In: Proceedings of the European Conference on Speech Communication and Technology, vol. 3, pp. 1211–1214 (1999)
56. Ling, Z.H., Wu, Y.J., Wang, Y.P., Qin, L., Wang, R.H.: USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In: Proc. the Blizzard Challenge Workshop (2006)
57. Lolive, D., Barbot, N., Boeffard, O.: Pitch and duration transformation with non-parallel data. Speech Prosody 2008 pp. 111–114 (2008)
58. Mariéthoz, J., Bengio, S.: Can a professional imitator fool a GMM-based speaker verification system? IDIAP Research Report 05-61 (2005)
59. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech., pp. 1895–1898 (1997)
60. Masuko, T., Hitotsumatsu, T., Tokuda, K., Kobayashi, T.: On the security of HMM-based speaker verification systems against imposture using synthetic speech. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech. (1999)
61. Masuko, T., Tokuda, K., Kobayashi, T., Imai, S.: Speech synthesis using HMMs with dynamic features. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP) (1996)
62. Masuko, T., Tokuda, K., Kobayashi, T., Imai, S.: Voice characteristics conversion for HMM-based speech synthesis system. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP) (1997)
63. Matrouf, D., Bonastre, J.F., Fredouille, C.: Effect of speech transformation on impostor acceptance. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), vol. 1, pp. I–I. IEEE (2006)
64. Matsui, T., Furui, S.: Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. Speech Commun. **17**(1-2), 109–116 (1995)
65. Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. **9**, 453–467 (1990)
66. Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B.: Transformation of formants for voice conversion using artificial neural networks. Speech communication **16**(2), 207–216 (1995)
67. Ogihara, A., Unno, H., Shiozakai, A.: Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification. IEICE transactions on fundamentals of electronics, communications and computer sciences **88**(1), 280–286 (2005)
68. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proceedings of Odyssey 2001: The Speaker and Language Recognition Workshop, pp. 213–218. Crete, Greece (2001)
69. Pellom, B.L., Hansen, J.H.: An experimental study of speaker verification sensitivity to computer voice-altered imposters. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), vol. 2, pp. 837–840. IEEE (1999)
70. Perrot, P., Aversano, G., Blouet, R., Charbit, M., Chollet, G.: Voice forgery using ALISP: indexation in a client memory. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), vol. 1, pp. 17–20. IEEE (2005)

71. Pilkington, N.C., Zen, H., Gales, M.J.: Gaussian process experts for voice conversion. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
72. Popa, V., Silen, H., Nurminen, J., Gabbouj, M.: Local linear transformation for voice conversion. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 4517–4520. IEEE (2012)
73. Quatieri, T.F.: Discrete-Time Speech Signal Processing Principles and Practice. Prentice-Hall, Inc. (2002)
74. Reynolds, D., Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing **3**, 72–83 (1995)
75. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing **10**(1), 19–41 (2000)
76. Saeidi, R., et al.: I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. Lyon, France (2013)
77. Saito, D., Watanabe, S., Nakamura, A., Minematsu, N.: Statistical voice conversion based on noisy channel model. Audio, Speech, and Language Processing, IEEE Trans. on **20**(6), 1784–1794 (2012)
78. Saito, D., Yamamoto, K., Minematsu, N., Hirose, K.: One-to-many voice conversion based on tensor representation of speaker space. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 653–656 (2011)
79. Satoh, T., Masuko, T., Kobayashi, T., Tokuda, K.: A robust speaker verification system against imposture using an HMM-based speech synthesis system. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech. (2001)
80. Shang, W., Stevenson, M.: Score normalization in playback attack detection. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 1678–1681. IEEE (2010)
81. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. Speech Communication **46**(3-4), 455–472 (2005)
82. Siddiq, S., Kinnunen, T., Vainio, M., Werner, S.: Intonational speaker verification: a study on parameters and performance under noisy conditions. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 4777–4780. Kyoto, Japan (2012)
83. Solomonoff, A., Campbell, W., Boardman, I.: Advances in channel compensation for SVM speaker recognition. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 629–632. Philadelphia, USA (2005)
84. Song, P., Bao, Y., Zhao, L., Zou, C.: Voice conversion using support vector regression. Electronics letters **47**(18), 1045–1046 (2011)
85. Stylianou, Y.: Voice transformation: a survey. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP), pp. 3585–3588. IEEE (2009)
86. Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion. Speech and Audio Processing, IEEE Trans. on **6**(2), 131–142 (1998)
87. 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. Acoust., Speech and Signal Process.(ICASSP), vol. 1, pp. I–I. IEEE (2006)
88. Sundermann, D., Ney, H.: VTLN-based voice conversion. In: Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on, pp. 556–559. IEEE (2003)
89. Toda, T., Black, A.W., Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. Audio, Speech, and Language Processing, IEEE Transactions on **15**(8), 2222–2235 (2007)
90. Villalba, J., Lleida, E.: Speaker verification performance degradation against spoofing and tampering attacks. In: FALA 10 workshop, pp. 131–134 (2010)
91. Villalba, J., Lleida, E.: Preventing replay attacks on speaker verification systems. In: Security Technology (ICCST), 2011 IEEE International Carnahan Conference on, pp. 1–8. IEEE (2011)

92. Wang, Z.F., Wei, G., He, Q.H.: Channel pattern noise based playback attack detection algorithm for speaker recognition. In: International Conferenceon Machine Learning and Cybernetics (ICMLC), vol. 4, pp. 1708–1713. IEEE (2011)
93. Woodland, P.C.: Speaker adaptation for continuous density HMMs: A review. In: Proc. ISCA Workshop on Adaptation Methods for Speech Recognition, p. 119 (2001)
94. Wu, C.H., Hsia, C.C., Liu, T.H., Wang, J.F.: Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. Audio, Speech, and Language Processing, IEEE Trans. on **14**(4), 1109–1116 (2006)
95. Wu, Z., Chng, E.S., Li, H.: Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. (2012)
96. Wu, Z., Chng, E.S., Li, H.: Conditional restricted boltzmann machine for voice conversion. In: the first IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE (2013)
97. Wu, Z., Kinnunen, T., Chng, E.S., Li, H.: Mixture of factor analyzers using priors from non-parallel speech for voice conversion. IEEE Signal Processing Letters **19**(12) (2012)
98. 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: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pp. 1–5. IEEE (2012)
99. Wu, Z., Larcher, A., Lee, K.A., Chng, E.S., Kinnunen, T., Li, H.: Vulnerability evaluation of speaker verication under voice conversion spoong: the effect of text constraints. In: Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc. Lyon, France (2013)
100. Wu, Z., Xiao, X., Chng, E.S., Li, H.: Synthetic speech detection using temporal modulation feature. In: Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.(ICASSP) (2013)
101. Wu, Z.Z., Kinnunen, T., Chng, E.S., Li, H.: Text-independent F0 transformation with non-parallel data for voice conversion. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
102. 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. Speech, Audio & Language Process. **17**(1), 66–83 (2009)
103. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. Eurospeech, ESCA Euro. Conf. on Speech Comm. and Tech., pp. 2347–2350 (1999)
104. Zen, H., Nankaku, Y., Tokuda, K.: Continuous stochastic feature mapping based on trajectory HMMs. Audio, Speech, and Language Processing, IEEE Transactions on **19**(2), 417–430 (2011)
105. Zen, H., Toda, T., Nakamura, M., Tokuda, K.: Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. **E90-D**(1), 325–333 (2007)
106. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication **51**(11), 1039–1064 (2009). DOI DOI:10.1016/j.specom.2009.04.004
107. Zetterholm, E., Blomberg, M., Elenius, D.: A comparison between human perception and a speaker verification system score of a voice imitation. In: Proc. of Tenth Australian International Conference on Speech Science & Technology, pp. 393–397. Macquarie University, Sydney, Australia (2004)