Spoofing and countermeasures for speaker verification: a survey

Zhizheng Wu[a,1,*], Nicholas Evans[b], Tomi Kinnunen[c], Junichi Yamagishi[d,e], Federico Alegre[b], Haizhou Li[a,f]

[a]*Nanyang Technological University, Singapore*
[b]*EURECOM, France*
[c]*University of Eastern Finland, Finland*
[d]*National Institute of Informatics, Japan*
[e]*University of Edinburgh, United Kingdom*
[f]*Institute for Infocomm Research, Singapore*

Abstract

While biometric authentication has advanced significantly in recent years, evidence shows the technology can be susceptible to malicious spoofing attacks. The research community has responded with dedicated countermeasures which aim to detect and deflect such attacks. Even if the literature shows that they can be effective, the problem is far from being solved; biometric systems remain vulnerable to spoofing. Despite a growing momentum to develop spoofing countermeasures for automatic speaker verification, now that the technology has matured sufficiently to support mass deployment in an array of diverse applications, greater effort will be needed in the future to ensure adequate protection against spoofing. This article provides a survey of past work and identifies priority research directions for the future. We summarise previous studies involving impersonation, replay, speech synthesis and voice conversion spoofing attacks and more recent efforts to develop dedicated countermeasures. The survey shows that future research should address the lack of standard datasets and the over-fitting of existing countermeasures to specific, known spoofing attacks.

Keywords: Automatic speaker verification, spoofing attack, countermeasure, security

Contents

1 Introduction **1**

2 Automatic speaker verification **2**
2.1 Feature extraction . . . 3
2.2 Speaker modeling and classification . . . 3
2.3 System fusion . . . 4

3 Vulnerability of speaker verification to spoofing **4**
3.1 Possible attack points . . . 4
3.2 Potential vulnerabilities . . . 4

4 Evaluation protocol **5**
4.1 Dataset design . . . 5
4.2 Evaluation metrics . . . 6

5 Spoofing and countermeasures **6**
5.1 Impersonation . . . 6
5.2 Replay . . . 8
5.3 Speech synthesis . . . 10
5.4 Voice conversion . . . 12

6 Discussion **14**
6.1 Spoofing . . . 14
6.2 Countermeasures . . . 15
6.3 Generalised countermeasures . . . 15

7 Issues for future research **15**
7.1 Large-scale standard datasets . . . 15
7.2 Evaluation metrics . . . 16
7.3 Open-source software packages . . . 17
7.4 Future directions . . . 18

8 Conclusions **18**

1. Introduction

Various distinctive and measurable physiological and behavioural traits have been investigated for biometric recognition (Jain et al., 2006). As our primary method of communication, speech is a particularly appealing modality. Individual differences in both physiological and behavioural characteristics, e.g. the vocal tract shape and intonation, can be captured and utilised for automatic speaker verification (ASV) (Kinnunen and Li, 2010).

Recent advances in channel and noise compensation techniques have significantly improved ASV performance to levels required for mass-market adoption. Reliable and efficient authentication is now possible in smartphone logical access scenarios (Lee et al., 2013) and in e-commerce (Nuance, 2013)

*Corresponding author
Email addresses: zhizheng.wu@ed.ac.uk (Zhizheng Wu), evans@eurecom.fr (Nicholas Evans), tkinnu@cs.uef.fi (Tomi Kinnunen), jyamagis@inf.ed.ac.uk (Junichi Yamagishi), alegre@eurecom.fr (Federico Alegre), hli@i2r.a-star.edu.sg (Haizhou Li)
[1]Now with the University of Edinburgh, United Kingdom

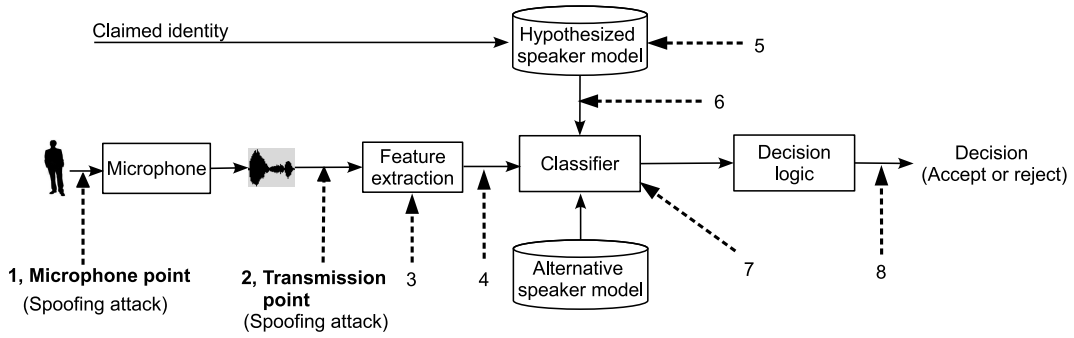


Figure 1: An illustration of a typical automatic speaker verification (ASV) system with eight possible attack points. Attacks at points 1-2 are considered as direct attacks whereas those at points 3-8 are indirect attacks.

for example. Even though ASV provides a low-cost and convenient approach to authentication, however, reliability in the face of spoofing remains a concern (Evans et al., 2013, 2014b).

A generic biometric system may be manipulated or attacked at various stages between sample acquisition and the delivery of an authentication result (Ratha et al., 2001; Faundez-Zanuy, 2004; Galbally et al., 2010). In the specific case of ASV as illustrated in Figure 1, attacks at both the microphone and transmission levels are generally considered to pose the greatest threat (Faundez-Zanuy et al., 2006). Here, an adversary, typically referred to as an impostor, might seek to deceive the system by impersonating another enrolled user at the microphone in order to manipulate the ASV result. Alternatively, captured speech signals can be intercepted and replaced at the transmission level by another specially crafted voice signal. Since speaker recognition is commonly used in telephony, or other unattended, distributed scenarios without human supervision or face-to-face contact, speech is arguably more prone to malicious interference or manipulation than other biometric signals; the potential for ASV systems to be spoofed is now well-recognised (Evans et al., 2013, 2014b; Wu and Li, 2013).

Prior to the consideration of spoofing, ASV systems were designed to distinguish between target speakers and zero-effort impostors. This research focuses on improving fundamental recognition performance, as opposed to security or robustness to spoofing and drove the community to investigate different approaches to speaker characterisation at the feature level including: (i) short-term spectral and voice source features, such as Mel-frequency cepstral coefficients (MFCCs) and glottal pulse features; (ii) prosodic and spectro-temporal features such as rhythm, pitch and other segmental information; (iii) high-level features such as phonetic, idiolect, and lexical features (Kinnunen and Li, 2010). Due to their simplicity and resulting ASV performance, most speaker verification systems utilise short-term spectral features. The literature shows that systems based on such features are vulnerable to spoofing; speech signals with corresponding features reflective of other speakers can be synthesised with ease (Evans et al., 2013, 2014b).

Numerous vulnerability studies suggest an urgent need to address spoofing. This can be accomplished via one of two general approaches. Some work, e.g. (Kinnunen et al., 2012) shows that advanced algorithms, such as joint factor analysis (Kenny, 2006), may offer an inherent protection from spoofing. The first approach is therefore to continue with the traditional pursuit of improved fundamental performance (i.e. in the face of only zero-effort impostors). The other approach involves the design of specific or generalised spoofing countermeasures. While both approaches will remain important, independent countermeasures have the advantage of being easily incorporated into existing ASV system and of being able to *detect* spoofing attempts. Research in this latter approach is in its relative infancy and greater attention will be needed in the future.

While the use of different datasets, protocols and metrics hinders such a task, this paper provides a survey of the past work. We compare the vulnerabilities of four different spoofing attacks considered thus far: impersonation, replay, speech synthesis and voice conversion. We then review anti-spoofing approaches or countermeasures for each form of attack. Finally we discuss directions for future work which will be necessary in order to address weaknesses in the current research methodology.

2. Automatic speaker verification

The task of an automatic speaker verification (ASV) system is to accept or reject a claimed identity based on a speech sample (Kinnunen and Li, 2010). There are two types of ASV systems: *text-dependent* and *text-independent*. Text-dependent systems assume fixed or prompted phrases which are usually the same for enrolment and for verification. Text-independent systems operate on arbitrary utterances, possibly spoken in different languages (Campbell Jr, 1997). Text-dependent ASV is generally better suited to authentication scenarios since higher recognition accuracy can then be achieved with shorter utterances. Nevertheless, text-independent systems also have utility, for example in call-centre applications including caller verification for telephone banking[2]. On account of evaluation sponsorship and dataset availability, text-independent ASV dominates the field and the research tends to place greater emphasis on surveillance applications rather than authentication.

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

This section describes briefly the state-of-the-art in ASV and the potential for the technology to be spoofed. More general and detailed overviews of the fundamentals (not specific to spoofing) can be found in (Campbell Jr, 1997; Bimbot et al., 2004; Kinnunen and Li, 2010; Li and Ma, 2010; Togneri and Pullella, 2011; Li et al., 2013)

2.1. Feature extraction

A speech signal has three-fold information: voice timbre, prosody and language content. Correspondingly, speaker individuality can be characterised by short-term spectral, prosodic and high-level idiolectal features. Short-term spectral features are extracted from short frames typically of 20-30 milliseconds duration. They describe the short-term spectral envelope which is an acoustic correlate of voice timbre. Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs) and perceptual linear prediction (PLP) are all popular spectral features.

Prosodic features are extracted from longer segments such as syllables and word-like units to characterise speaking style and intonation. These features, such as pitch, energy and duration, are less sensitive to channel effects. However, due to their sparsity, the extraction of prosodic features requires relatively large amounts of training data (Adami et al., 2003; Kajarekar et al., 2003; Shriberg et al., 2005), and pitch extraction algorithms are generally unreliable in noisy environments (Gerhard, 2003).

High-level features (Doddington, 2001; Reynolds et al., 2003) are extracted from a lexicon (or other discrete tokens) to represent speaker behaviour or lexical cues. High-level features are considered to be even less sensitive to channel and noise effects than spectral and prosodic features. However, the extraction of high-level features requires considerably more complex front-ends, such as those which employ automatic speech recognition (Kinnunen and Li, 2010; Li and Ma, 2010).

2.2. Speaker modeling and classification

Approaches to text-independent ASV generally focus on modelling the feature distribution of a target speaker. The theoretical framework of most ASV systems involves the computation of a *log-likelihood ratio* (LLR) score,

$$\ell = \log \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_1)}, \tag{1}$$

and its comparison to a pre-determined threshold in order to decide in favour of either the target hypothesis H_0 (same speaker) or the alternative hypothesis H_1 (different speaker). Here $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$ is a sequence of feature vectors while $p(\mathbf{X}|H_0)$ and $p(\mathbf{X}|H_1)$ denote the likelihood of each hypothesis. Intuitively, the alternative model $p(\mathbf{X}|H_1)$ helps to normalise common effects not related to speaker identity. There are many different ways to implement Eq. (1). In the classical approach (Reynolds and Rose, 1995), features $\mathbf{X}$ are typically MFCCs and the acoustic models are Gaussian mixture models (GMMs) (see below). With more modern techniques, $\mathbf{X}$ can also be high-dimensional i-vectors (Dehak et al., 2011) modelled with probabilistic linear discriminant analysis (PLDA) back-ends (Li et al., 2012) (see below). Even so, GMMs are still needed for i-vector extraction and thus we provide a more detailed presentation of the GMM in the following.

GMMs have been used intensively and their combination with a universal background model (UBM) has become the *de facto* standard, commonly referred to as the GMM-UBM approach (Reynolds et al., 2000). Here, speech samples pooled from a large number of speakers are used to estimate a speaker-independent UBM using a maximum likelihood (ML) criterion; the UBM likelihood corresponds to $p(\mathbf{X}|H_1)$ in Eq. (1). Speaker-dependent models, used in determining $p(\mathbf{X}|H_0)$ in Eq. (1), are then derived from the UBM with maximum a posteriori (MAP) adaptation using the speech samples of a target speaker (Gauvain and Lee, 1994). The target speaker and UBM models are used as the hypothesised and alternative speaker models respectively.

As a two-class classification problem, the discrimination between hypothesised and alternative speaker models is key to performance. The combination of GMM *supervectors* and support vector machine (SVM) classifiers was developed to enable the discriminative training of generative models (Campbell et al., 2006). This idea lead to the development of many successful model normalisation techniques including nuisance attribute projection (NAP) (Solomonoff et al., 2005; Burget et al., 2007) and within-class covariance normalisation (WCCN) (Hatch et al., 2006). These techniques all aim to compensate for intersession variation and channel mismatch.

Parallel to the development of SVM-based discriminative approaches, generative factor analysis models were pioneered in (Kenny, 2006; Kenny et al., 2007, 2008). In particular, *joint factor analysis* (JFA) (Kenny, 2006) can improve ASV performance by incorporating distinct speaker and channel subspace models. These subspace models involve the training of various hyper-parameters and generally require large quantities of labeled utterances. JFA subsequently evolved into a simplified *total variability model* or 'i-vector' approach which is now the state of the art (Dehak et al., 2011). An i-vector represents an arbitrary utterance, encoded via its GMM mean supervector, with a low dimensional vector of latent variables. From this perspective, i-vector extraction is a dimensionality reduction process, which accordingly supports the application of traditional pattern recognition techniques to i-vector modelling and comparison. *Probabilistic linear discriminant analysis* (PLDA) (Li et al., 2012), a factor analysis technique originally developed for face recognition (Prince and Elder, 2007), is the most popular approach. The normalisation of i-vectors to lie on a unit sphere is also popular as a pre-processing technique for the PLDA back-end (Garcia-Romero and Espy-Wilson, 2011).

In contrast to text-independent systems, text-dependent systems not only model the feature distribution, but also the language content. The underlying feature extraction, speaker modelling and classification approaches developed for text-independent systems, including i-vector and PLDA models, can also be applied within text-dependent systems with minor modifications (Larcher et al., 2013b; Stafylakis et al., 2013).

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*. The motivation is based on the assumption that multiple, independently trained recognisers together capture different aspects of the speech signal not covered by a single classifier alone. Fusion also provides a convenient vehicle for large-scale research collaborations promoting independent classifier development and benchmarking (Saeidi et al., 2013).

Different classifiers can involve different features, classifiers, or hyper-parameter training sets (Brümmer et al., 2007; Hautamäki et al., 2013b). 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.

3. Vulnerability of speaker verification to spoofing

3.1. Possible attack points

A typical ASV system involves two processes: offline enrolment and runtime verification. During the offline enrolment, a target speaker model is trained using features extracted from a sample of speech. The runtime verification process is illustrated in Figure 1, where a speaker first asserts an identity claim and then provides a sample of his/her speech. Features similarly extracted from this sample are compared to the model in order to determine whether or not the speaker matches the claimed identity.

In practice the sample is compared to two models, one corresponding to the hypothesised speaker and a second representing the alternative hypothesis. The classifier determines a match score which represents the relative similarity of the sample to each of the two models. Finally, the decision logic module uses the relative score (usually, a log-likelihood ratio) to either accept or reject the identity claim.

These components and the links between them all represent possible attack points (Ratha et al., 2001). Eight such vulnerability points for a generic ASV system are also illustrated in Figure 1. They can be categorised as follows:

Direct attacks, also referred to as *spoofing attacks*, can be applied at the microphone level as well as the transmission level – labelled as attack points 1 and 2 in Figure 1. Examples include the impersonation of another person or the presentation of a pre-recorded or synthesised speech signal at the microphone.

Indirect attacks are performed within the ASV system itself – labelled as attack points 3 to 8 in Figure 1. Indirect attacks generally require system-level access, for example attacks which interfere with feature extraction (points 3 and 4), models (points 5 and 6) or score and decision logic computation (points 7 and 8).

Even if for some physical or logical access scenarios, attacks may be applied only at the microphone (Lee et al., 2013), and in contrast to the wider literature pertaining to other biometric modalities (Ratha et al., 2001), we include transmission level attacks (point 2 in Figure 1) as a form of direct attack in the context of ASV. This is justified on account of the often-distributed nature of ASV systems which might allow for an attacker to interfere with the microphone signal. There is also potential for spoofed speech signals can be injected immediately prior to transmission while bypassing the microphone entirely. To exemplify, the Skype Voice Changer[3] allows a voice signal to be manipulated after capture but prior to transmission.

Since neither microphone level nor transmission level attacks necessarily require system-level access, they are the most easily implemented attacks and are thus the greatest threat to typical ASV systems (Faundez-Zanuy et al., 2006). They are accordingly the focus in the remainder of this paper. In past studies of ASV spoofing, impersonation and replay attacks are assumed to apply at the microphone. Even if speech synthesis and voice conversion attacks may also be applied at the microphone, in the literature they generally target the transmission level, thereby bypassing the microphone.

3.2. Potential vulnerabilities

This section explains the potential for typical ASV systems to be spoofed. We focus on two key ASV modules: feature extraction and speaker modelling.

3.2.1. Feature extraction

All three feature representations described in Section 2.1 are potentially vulnerable to spoofing attacks. Due to their simplicity and performance, short-term spectral features are the most popular. Ignoring any channel effects, replay attacks which use a pre-recorded speech sample can faithfully reflect the spectral attributes of the original speaker. State-of-the-art speech synthesisers contain models of short-term spectral characteristics and can thus be adapted to reflect those of a specific, target speaker (Ling et al., 2012). Voice conversion can also generate speech signals whose spectral envelope reflects that of a target speaker (Matrouf et al., 2006). Figure 2 illustrates the effect of voice conversion on an impostor speech signal (dashed blue profile). The spectral envelope corresponding to a single speech frame is shifted towards that of a given, target speaker (green profile). ASV systems which use short-term spectral features are thus vulnerable to spoofing.

Prosodic characteristics may also be mimicked through impersonation and appropriately trained speech synthesis and voice conversion systems. For example, some speech synthesisers can generate fundamental frequency trajectories which are highly correlated with those of a given, target speaker (Qian et al., 2011).

High-level features reflect language content and speaker behaviour, e.g. the choice of words. Although such features might be useful for speaker characterisation, they may be mimicked relatively easily. For example, replay attacks are performed using a target speaker's pre-recorded speech, which will naturally

[3] `http://www.skypevoicechanger.com/`

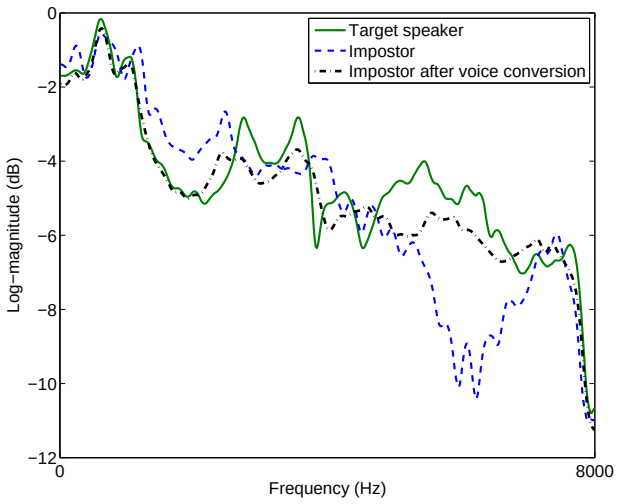


Figure 2: An illustration of voice conversion and the potential for spoofing. The spectral envelope of an impostor's speech signal is shifted towards that of a given target speaker.

contain the same or similar language content and speaker behaviour. Speech signals with similar language content can also be synthesised with ease.

3.2.2. Speaker modeling

Most approaches to speaker modelling, be they applied to text-independent or text-dependent ASV, have their roots in the standard GMM. Most lack the modelling of temporal sequence information, a key characteristic of human speech, which might otherwise afford some protection from spoofing; most models of the feature distributions used in typical speech synthesis and voice conversion algorithms assume independent features of observations, but are nonetheless effective as spoofing attacks. As shown in (Kons and Aronowitz, 2013), HMM-based systems, which capture temporal information, are more robust to spoofing than GMM-based systems when subject to the same spoofing attack.

While preliminary studies of fused ASV system approaches to anti-spoofing were reported in (Riera et al., 2012), some insight into their likely full potential can be gained from related work in fused, multi-modal biometric systems. A long-lived 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 (Rodriques et al., 2009; Akhtar et al., 2012) suggests it might suffice to spoof only *one* modality (or sub-system) under a score fusion setting in the case where the spoofing of a single, significantly weighted sub-system is particularly effective. Thus, traditional fusion techniques may not provide significantly increased robustness to spoofing unless they are coupled with dedicated spoofing countermeasures.

4. Evaluation protocol

Here we present a generic experimental protocol which applies to the majority of past work. We discuss database design and evaluation metrics with a focus on the comparability of baseline results with those of vulnerability and countermeasure studies.

4.1. Dataset design

While past studies of spoofing have used a range of different datasets (Alegre et al., 2014) there are some similarities in the experimental protocols. Essential to them all is the meaningful comparison of baseline performance to that for the same system when subjected to spoofing attacks. The majority of the past spoofing studies reported in this paper conform to the general assessment framework illustrated in Figure 3. The diagram illustrates three possible inputs: genuine, zero-effort impostor and spoofed speech.

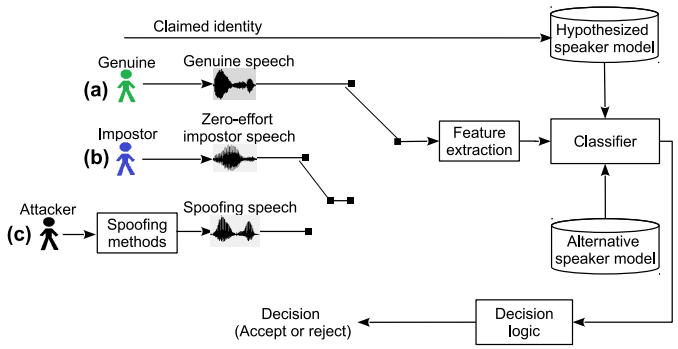


Figure 3: Illustration of the general framework used in past spoofing attack studies. There are three kinds of input: (a) genuine speech; (b) zero-effort impostor speech; and (c) spoofed speech. An evaluation using (a) and (b) relates to a standard, baseline ASV evaluation, whereas an evaluation using (a) and (c) is used to evaluate spoofing and countermeasure performance. Generally, (c) represents spoofed version of (b), and thus (b) has the same number of trials as (c).

The combination of genuine and zero-effort impostor tests comprise a standard, baseline ASV evaluation. In this case we suppose the protocols for such an evaluation stipulate M target trials and N impostor trials. A new dataset suitable for the study of spoofing is derived from the baseline by replacing all impostor trials with spoofed trials. For example, in a study of spoofing through voice conversion, the speech data of all impostor trials are converted towards the target client in order to generate new speech data for spoofing evaluations. There are then N spoofing trials which replace the N previous impostor trials.

Referring once again to Figure 3, baseline performance is assessed using the pool of M genuine trials (a) and N impostor trials (b), while that under spoofing is assessed with the pool of M genuine trials (a) and N spoofing trials (c). If the ASV system used for both baseline and spoofing tests is the same, then scores and decisions for all genuine trials will remain unchanged. The baseline performance and that under spoofing is

thus directly comparable and the difference between them reflects the vulnerability of the system to the particular spoofing attack considered.

4.2. Evaluation metrics

The evaluation of ASV systems requires large numbers of two distinct tests: target tests, where the speaker matches the claimed identity, and impostor tests, where the identities differ. Accordingly, the ASV system is required to either accept or reject the identity claim, thereby resulting in one of four possible outcomes, as illustrated in Table 1. There are two possible correct outcomes and two possible incorrect outcomes, namely false acceptance (or false alarm) and false rejection (or miss). Statistics acquired from many independent tests (trials) are used to estimate the false acceptance rate (FAR) and the false rejection rate (FRR). The FAR and FRR are complementary in the sense that, for a variable threshold and otherwise fixed system, one can only be reduced at the expense of increasing the other. In practice, all system parameters are optimised to minimise the balance between FAR and FRR, which is commonly measured in terms of the equal error rate (EER)[4], although this is certainly not the only optimisation criterion.

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

	Decision	
	Accept	Reject
Genuine	Correct acceptance	False rejection
Impostor	False acceptance	Correct rejection

In a spoofing scenario, an attacker attempts to bias the system outcome towards accepting a false identity claim. Equivalently, spoofing attacks will increase the FAR for a fixed decision threshold optimised on the standard baseline ASV dataset. Increases in the FAR (for a fixed FRR) are also reflected in the EER. As is common in the literature, both metrics may thus be used to gauge the robustness of an ASV system to spoofing.

To prevent spoofing attacks, countermeasures have been developed to decide whether a particular trial is a licit access attempt or a spoofing attack. Ideally, countermeasures should decrease the FAR in the event of spoofing attacks while not increasing the FRR in the case of genuine access attempts. Nonetheless, similar to the decisions of a regular ASV system as illustrated in Figure 4, a practical, stand-alone countermeasure will inevitably lead to some false acceptances, where a spoofing attack remains undetected, in addition to false rejections, where genuine attempts are identified as spoofing attacks.

In addition to EER, FAR and FRR metrics, the detection cost function (DCF) is also popular. The DCF represents a trade-off between the FAR and FRR using a priori probabilities of target and non-target events. Although the DCF has been used widely for the evaluation of ASV performance, it has not been used extensively in the spoofing and countermeasure literature. Accordingly, in the following sections we report results only in terms of EERs, FARs and FRRs.

[4]EER corresponds to the operating point at which FAR=FRR.

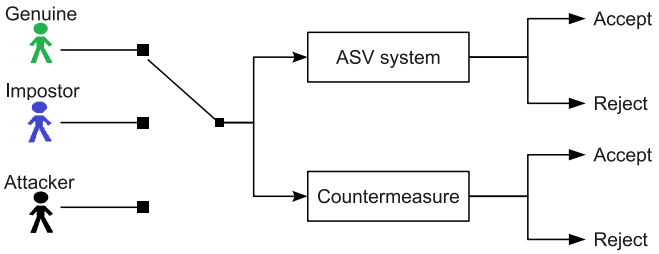


Figure 4: An illustration of decisions taken by a stand-alone ASV system and countermeasure. A stand-alone ASV system could falsely accept an impostor or a spoofed trial (a false acceptance), while a stand-alone countermeasure could reject a genuine trial (a false rejection).

5. Spoofing and countermeasures

This section reviews past work to evaluate the vulnerability of typical ASV systems to spoofing and parallel efforts to develop anti-spoofing countermeasures. Spoofing implies an attack at either the microphone or transmission level using a manipulated or synthesised speech sample in order to bias the system towards verifying a claimed identity. We consider impersonation, replay, speech synthesis and voice conversion while concentrating on three different aspects: (i) the practicality of each spoofing attack; (ii) the vulnerability of ASV systems when subjected to such attacks, and (iii) the design of a realistic datasets for experimentation. With regard to countermeasures, we focus on: (i) the effectiveness of a countermeasure in preventing specific spoofing attacks, and (ii) the generalisation of countermeasures in protecting against varying attacks.

5.1. Impersonation

Impersonation is one of the most obvious approaches to spoofing and refers to attacks using human-altered voices, otherwise referred to as human mimicking. Here, an attacker tries to mimic a target speaker's voice timbre and prosody without computer-aided technologies.

5.1.1. Spoofing

The work in (Lau et al., 2004) showed that non-professional impersonators can readily adapt their voice to overcome ASV, but only when their natural voice is already similar to that of the target speaker (closest targets were selected from the YOHO corpus using a speaker recognition system). Further work in (Lau et al., 2005) 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 markedly different to that of the impersonator. However, experiments reported in (Mariéthoz and Bengio, 2006) suggest that, while professional impersonators are more effective than the untrained, even they are *unable* to consistently spoof an ASV system. A more recent study (Hautamäki et al., 2013a) analysed the vulnerability of both classical GMM-UBM and state-of-the-art i-vector systems to impersonation attacks. In this study, five Finnish public figures were used as target speakers, all of whom were impersonated by a professional impersonator. Similar to the

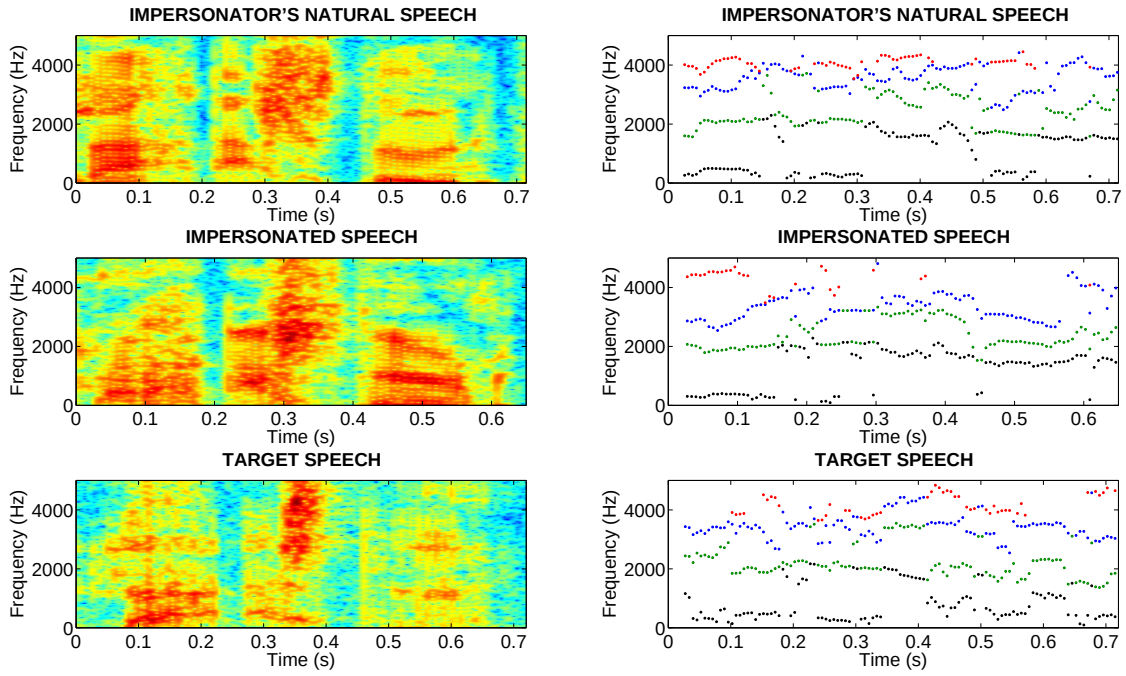


Figure 5: An example of speech impersonation. Finnish target speaker pronounces "…lehdistöön" ("…to the press…"), a chunk extracted from a long interview passage. The impersonator attempts to sound like the target. Spectrogram and formant tracks (F1 through F4) of the impersonator's own voice (top), impersonation (middle) and the target speaker (bottom). The spectrograms and formants (Burg's method) were computed with Praat (Boersma and Weenink, 2014) from material collected by (Leskelä, 2011) and used in (Hautamäki et al., 2013a). The target speaker is the current president of Finland, Sauli Niinistö. Comparing the top and middle figures, the impersonator can modify his voice away from his natural vocal tract configuration (for instance, F3 is generally lowered). Nevertheless, the formants do not quite match those of the target speaker. Perceptually, the impersonation sounds convincing to a native listener.

findings in (Mariéthoz and Bengio, 2006), the impersonator was unable to spoof either ASV system.

In addition to specific spoofing assessments, some insights into potential vulnerabilities can be drawn from various acoustic-phonetic studies of impersonation (Endres et al., 1971; Blomberg et al., 2004; Eriksson and Wretling, 1997; Zetterholm et al., 2004; Farrús et al., 2008; Amin et al., 2014). An example of impersonation, in terms of spectrogram and formants, is illustrated in Figure 5. The acoustic-phonetic studies show that, while imitators tend to be effective in mimicking long-term prosodic F0 patterns and speaking rates, they may be less effective in mimicking formant and other spectral characteristics. For instance, the imitator involved in the studies reported in (Eriksson and Wretling, 1997) was not successful in translating his formant frequencies towards the target, whereas different findings are reported in (Kitamura, 2008).

An interesting recent study (Amin et al., 2014) involves disguised speech material from three professional voice-over artists producing 27 distinct voice identities — interestingly, without any pre-specified target speakers, giving the impersonators artistic freedom in making up some virtual voice identities. One of the key observations was that the change in the vocal space (measured through F1 and F2) under impersonation cannot be described by a simple global transform; formant changes are vowel-specific. The same study also investigated glottal parameters (open quotient) measured from parallel electro-glottographic (EEG) recordings and found that impersonators have an active voluntary control over their vocal fold patterns. Unsurprisingly, the impersonators varied the mean and standard deviation of both F0 and the speaking rate to create distinct speaker identities.

Characteristic to all studies involving professional impersonators 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. Overall, these aspects make 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 (Perrot et al., 2005; Hautamäki et al., 2014).

Even if the statistical evidence from impersonation studies is limited, and the conclusions are somewhat inconsistent, there is alternative evidence for the potential of impersonation attacks. As discussed in (Campbell Jr, 1997; Doddington et al., 1998) some impostor speakers have natural potential to be confused with other speakers. Similarly, certain target speakers may be more easily impersonated than others. The work in (Stoll and Doddington, 2010) demonstrated the existence of such speakers in the NIST 2008 corpus and their effect on a wide range of modern ASV systems. These observations are not specific

Table 2: Summary of impersonation spoofing attack studies. ASV=automatic speaker verification, FAR=false acceptance rate, IER = identification error rate and k-NN = k-nearest neighbour. Note that IER is not comparable with FAR.

Study	# target speaker	# impersonators	FAR or IER			
			ASV system	Feature	Before spoofing	After spoofing
(Lau et al., 2004)	6	2	GMM-UBM	MFCCs	≈ 0 %	30 % ~ 35 %
(Lau et al., 2005)	4	6	GMM-UBM	MFCCs	≈ 0 %	10 % ~ 60 %
(Farrús et al., 2010)	5	2	k-NN	Prosodic features	5 % (IER)	22 % (IER)
(Hautamäki et al., 2013a)	5	1	i-vector	MFCCs	9.03 %	11.61 %

to ASV and similar findings have been reported in the general biometric literature (Yager and Dunstone, 2010).

In all cases, so-called *wolves* and *lambs* (Campbell Jr, 1997; Doddington et al., 1998) leave systems vulnerable to spoofing through the careful selection of target identities. Conversely, in order to impersonate a given individual, crowd-sourcing may be used to identify an impostor whose natural voice is similar to that of the target (Panjwani and Prakash, 2014). The work in (Lau et al., 2005) and (Stoll and Doddington, 2010) showed how ASV systems themselves or even acoustic features alone may be employed to identify 'similar' speakers in order to provoke false acceptances.

Past studies involving impersonation attacks are summarised in Table 2. It shows a degree of inconsistency in their finding with various ASV systems and feature representations. In addition, all four studies were conducted with datasets containing only a small number of speakers. In general, further studies will be needed to fully understand the effectiveness of impersonation.

5.1.2. Countermeasures

While the threat of impersonation is not fully understood it is perhaps not surprising that there is virtually 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 artefacts from synthesis or conversion systems, impersonators are live human beings who produce entirely natural speech. Interestingly, some related work (Amin et al., 2013, 2014) has addressed the problem of *disguise* detection[5]. The rationale behind the disguise detector developed in (Amin et al., 2013, 2014) is that impersonators are less practised with the impersonated voices and consequently exhibit larger (exaggerated) acoustic parameter variation under disguise. Specifically, the disguise detectors in (Amin et al., 2013, 2014) used quadratic discriminant on the first two formants to quantify the amount of acoustic variation on a vowel-by-vowel basis. Despite promising disguise detection results – 95.8 % to 100.0 % in (Amin et al., 2013) – the method requires vowel segmentation which was implemented through forced-alignment followed by manual correction.

[5]While the spoofing attacks discussed in this article are meant to increase false acceptance rate, disguise is the opposite problem where one wishes to be *not* recognized as herself, thereby increasing false rejection (miss) rate.

It might be beneficial to investigate new metrics to predict how easy or difficult it might be to impersonate a certain target (Stoll and Doddington, 2010), and then to develop specific fall-back mechanisms to cope with such speakers during runtime recognition.

5.2. Replay

Replay is a form of spoofing whereby an adversary attacks an ASV system using a pre-recorded speech sample collected from a genuine target speaker. The speech sample can be any recording captured surreptitiously and even concatenated speech samples extracted from a number of shorter segments, for example to overcome text-dependent ASV systems (Villalba and Lleida, 2011b). Replay is a simple spoofing attack, requiring no specific knowledge in speech processing. In addition, due to the availability of high quality and low-cost recording devices, such as smart phones, replay spoofing attack are arguably the most accessible and therefore present a significant threat. An example of a practical replay attack is presented in Figure 6. Here, a smart phone is used to replay a pre-recorded speech sample in order to unlock another smart phone which uses speaker verification technology for logical access authentication.

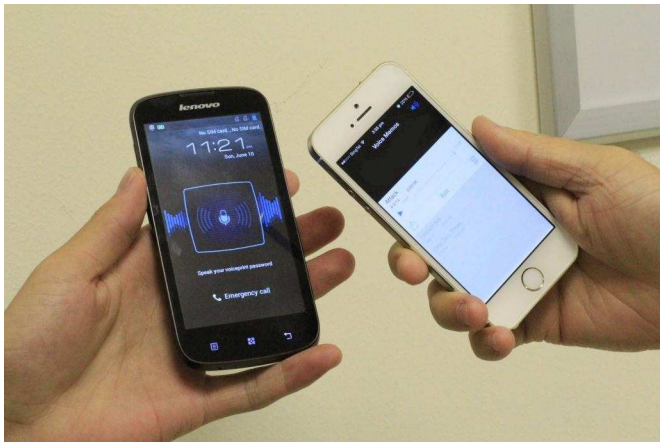


Figure 6: An example of replay attack in practical situation. The left phone (black color) is the smart phone with a voice-unlock function for user authentication as reported in (Lee et al., 2013). The right phone (white color) is used to replay a pre-recorded speech sample to unlock the left phone.

5.2.1. Spoofing

Even though they are among the most simple and easily implemented, only a small number of studies have addressed replay attacks. In those thus far reported, attacks are generally

Table 3: A summary of different studies involving replay spoofing attacks. CMs = Countermeasures.

Study	# target speaker	ASV system	Before spoofing	After spoofing		With CMs	
			EER/FAR	EER	FAR	EER	FAR
(Lindberg et al., 1999)	2	Text-Dependent HMM	1.1 ~ 5.6 %	27.3 ~ 70.0 %	89.5 ~ 100 %	n/a	n/a
(Villalba and Lleida, 2011a)	5	JFA	0.71 %	≈20 %	68.00 %	0 ~ 14 %	0 ~ 17 %
(Wang et al., 2011)	13	GMM-UBM	n/a	40.17 %	n/a	10.26 %	n/a

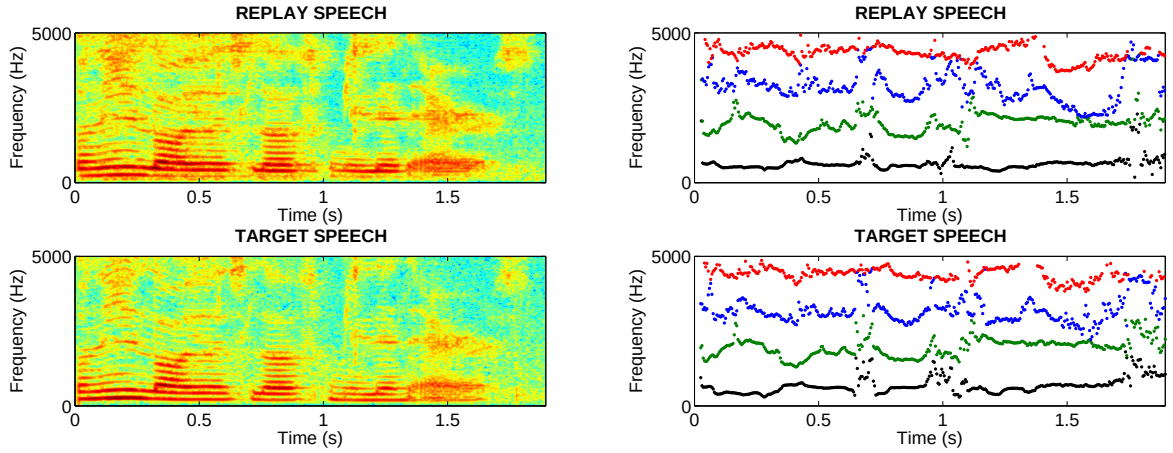


Figure 7: An example of replay. An English target speaker pronounces "only lawyers love millionaires". The original data is from RSR2015 (Larcher et al., 2012, 2014). The attacker uses a smartphone to record the genuine speech, and then replays it using a laptop, which is also used to record the playback speech. Spectrogram and formant tracks (F1 through F4) of the replay speech (top) and the genuine speech (botton) were computed using Praat (Boersma and Weenink, 2014). It is clearly observed that the spectrogram and formant tracks of the replay speech are almost indistinguishable from the genuine speech.

assumed to occur at the microphone level, although they can also be implemented at the transmission level in the case that replayed speech signals are injected immediately prior to transmission.

Vulnerabilities to replay attack were first evaluated in (Lindberg et al., 1999). The threat was assessed in the context of a hidden Markov model (HMM) text-dependent, digit sequence ASV system with attacks constructed from the concatenation of pre-recorded, isolated digits. Replay attacks were shown to provoke a significant increase in both the EERs and FARs. In particular, the EERs increased from 1.1 % and 5.6 % to 27.3 % and 70.0 % for male and female speakers, respectively. With the same threshold, the FARs were shown to increase from 1.1% and 5.6 % to 89.5 % and 100 % for male and female speakers, respectively. The significant variation between male and female speakers lies in the use of only a single speaker in each case.

Vulnerabilities in the context of text-independent ASV were assessed in (Villalba and Lleida, 2010) and (Villalba and Lleida, 2011a). Both studies used pre-recorded speech samples which were collected using a far-field microphone and then replayed in a mobile telephony scenario. Results showed that the FAR of a joint factor analysis (JFA) system increased from 0.71 % to almost 68 % as a result of replay attacks. Both studies involved only five speakers.

A physical access scenario was considered in (Wang et al., 2011). Although no baseline statistics were reported, a text-independent GMM-UBM system was shown to give an EER of 40.17 % when subjected to replay attacks. This study used a dataset collected from 13 speakers.

Figure 7 presents an example of replay speech in comparison to the genuine speech. It shows that the spectrogram and formant trajectories of the replay speech (upper images) have a highly similarity to those of the genuine speech (lower images). We can infer that the spectral features extracted from such a spectrogram will match the feature distribution of the target speaker to a considerable degree. Thus, it is easy to understand that ASV systems using spectral features are vulnerable to replay attacks.

A summary of the work involving replay spoofing attacks is presented in Table 3. Even if they are all based on a small number of speakers, all three studies are consistent in their findings: no matter what the ASV system, replay attacks provoke significant increases in FARs.

5.2.2. Countermeasures

Recently, due to the mass-market adoption of ASV techniques (Lee et al., 2013; Nuance, 2013) and the awareness and simplicity of replay attacks, both industry (Nuance, 2013) and academia (Shang and Stevenson, 2010; Villalba and Lleida, 2011a,b; Wang et al., 2011) have shown an interest in developing replay attack countermeasures.

The first approach to replay detection was reported in (Shang and Stevenson, 2010) in the context of a text-dependent ASV system using fixed pass-phrases. The detector is based upon the comparison of new access samples with stored instances of past access attempts. New accesses are identified as replay attacks if

they produce a similarity score higher than a pre-defined threshold. Detection performance was assessed using a database of genuine and replayed accesses collected across three different communication channels and using three different replay devices. A large number of experiments confirmed that the detector succeeded in lowering the EER in most of the playback detection experiments conducted.

An alternative countermeasure based upon spectral ratio and modulation indexes was proposed in (Villalba and Lleida, 2011a,b). The motivation stems from the increase in noise and reverberation which occurs as a result of replaying far-field recordings. The spectrum is flattened as a result and thus the modulation index is reduced. A support vector machine was used to model the spectral and modulation indexes of genuine and replayed recordings collected across both landline and GSM telephone channels. The countermeasures were shown to reduce the FAR of a text-independent joint factor analysis (JFA) ASV system from 68 % to 0 % and 17 % for landline and GSM channels, respectively.

A replay attack countermeasure based on the detection of channel noise was proposed in (Wang et al., 2011). Licit recordings only contain channel noise from the recording device of the ASV system, while replay attacks incur additional channel noise introduced by both the recording device and the loudspeaker used for replay. Thus, the detection of channel effects beyond those introduced by the recording device of the ASV system serves as an indicator of replay attack. Experiments showed that the performance of a baseline GMM-UBM system with a EER of 40.17 % under spoofing fell to 10.26 % with the countermeasure.

While related to a multimodal scenario with both speaker and face recognition,(Bredin et al., 2006) proposed a replay attack detection algorithm based on the lack in correspondence between acoustic and visual signals. Under replay attack an error rate of 0 % was achieved when the visual signal consisted only of a still photo.

The performance of ASV systems with replay attack countermeasures is summarised in Table 3. Even if all the example studies involve only a small number of speakers, it is clear that replay attacks provoke significant increases in the reported FARs. While countermeasures are generally effective in reducing the FARs, they remain significantly higher than those of the respective baselines. Further work is thus required to develop more effective countermeasures.

5.3. 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 (Klatt, 1980). 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) and concatenated them according to the given phoneme sequence by applying signal processing, such as linear predictive (LP) analysis, to the units (Moulines and Charpentier, 1990). 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, some commercial (Hunt and Black, 1996; Breen and Jackson, 1998; Donovan and Eide, 1998; Beutnagel et al., 1999; Coorman et al., 2000). In the late 1990s another data-driven approach emerged. 'Statistical parametric speech synthesis' has grown in popularity in recent years (Yoshimura et al., 1999; Ling et al., 2006; Black, 2006; Zen et al., 2007). 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 parameters and excitation parameters in order to generate a speech waveform. In addition to the four major approaches, inspired by advances in deep neural network (DNN)-based speech recognition (Hinton et al., 2012), new data-driven, DNN-based approaches have also been actively investigated (Zen et al., 2013; Ling et al., 2013; Lu et al., 2013; Qian et al., 2014).

The first three approaches are unlikely to be effective in ASV spoofing. The first two approaches do not provide for the synthesis of speaker-specific formant characteristics, whereas 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 (Zen et al., 2009; Yamagishi et al., 2009) can learn 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) (Leggetter and Woodland, 1995; Woodland, 2001).

Table 4: A summary of speech synthesis spoofing attack studies. CMs = Countermeasures. Note that all the studies listed in this Table do not provide EERs of the ASV systems after spoofing attacks.

Study	# target speaker	ASV system	FAR		
			Before spoofing	After spoofing	With CMs
(Lindberg et al., 1999)	2	HMM	5.6 %	38.9 %	n/a
(Masuko et al., 1999)	20	HMM	0.00 %	70 %	n/a
(De Leon et al., 2012a)	283	GMM-UBM	0.28 %	86 %	2.5 %
(De Leon et al., 2012a)	283	SVM	0.00 %	81 %	2.5 %

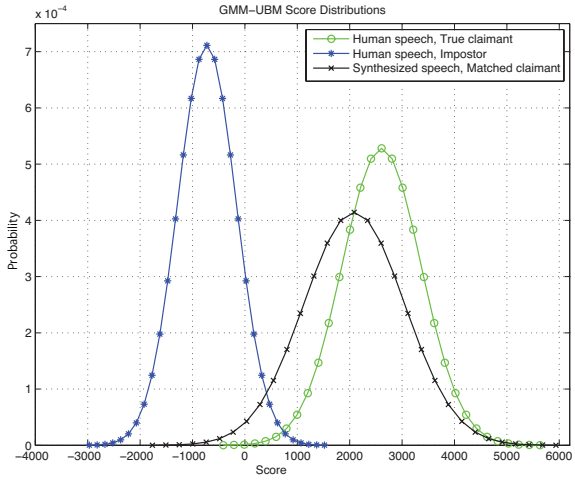
5.3.1. Spoofing

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 (Lindberg et al., 1999; Foomany et al., 2009; Villalba and Lleida, 2010). Although speech synthesis attacks can also be applied at the microphone level, the majority of past work assumes attacks at the transmission level.

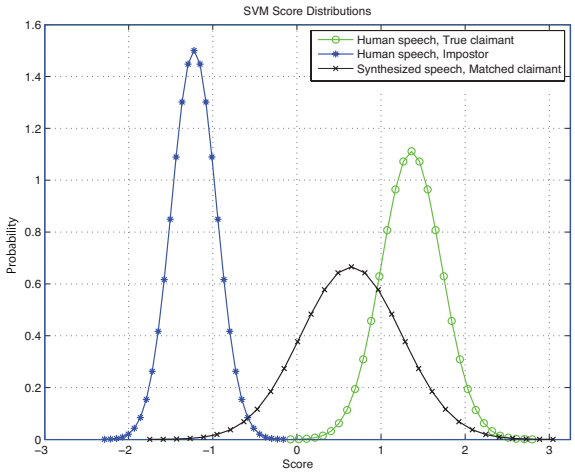
ASV vulnerabilities to HMM-based synthetic speech were first demonstrated over a decade ago (Masuko et al., 1999, 2000). The work used acoustic models adapted to specific human speakers (Masuko et al., 1996, 1997) and was performed using an HMM-based, text-prompted ASV system (Matsui and Furui, 1995). Feature vectors were scored against speaker and background models composed of concatenated phoneme models. When tested with genuine 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 %. This work involved only 20 speakers.

Larger scale experiments using the Wall Street Journal corpus containing in the order of 283 speakers and two different ASV systems (GMM-UBM and SVM using Gaussian supervectors) were reported in (De Leon et al., 2010b,a, 2012a). Using a state-of-the-art HMM-based speech synthesiser, the FAR was shown to rise from 0.28 % and 0 % to 86 % and 81 % for the GMM-UBM and SVM systems respectively. This result is due to the significant overlap in the distribution of ASV scores for genuine and synthetic speech, as shown in Figure 8. Spoofing experiments using HMM-based synthetic speech and a commercial, forensic speaker verification tool were also reported in (Galou, 2011) and reached similar findings.

Table 4 summarises the past work on speech synthesis attacks. In contrast to studies of impersonation and replay attacks, those of speech synthesis attacks generally use relatively large-scale datasets, usually of high-quality clean speech. All the past work confirms that speech synthesis attacks are able to increase significantly the FAR of all tested ASV systems, including those at the state of the art.



(a) GMM-UBM ASV System



(b) SVM using GMM supervectors ASV system

Figure 8: Approximate score distributions for (a) GMM-UBM and (b) SVM using GMM supervector SV systems with human and synthesized speech reported in (De Leon et al., 2012a). Distributions for human speech, true claimant (green lines, o) and synthesized speech, matched claimant (black lines, x) have significant overlap leading to a 81 % acceptance rate for synthetic speech with matched claims.

5.3.2. Countermeasures

Most approaches to detect synthetic speech rely on processing artefacts specific to a particular synthesis algorithm. Based on the observation that the dynamic variation in the speech parameters of synthetic speech tend to be less than those of natural speech, (Satoh et al., 2001) investigated the use of intra-frame differences as a discriminative feature. This method works well in detecting HMM-based synthetic speech without global variance compensation (Tomoki and Tokuda, 2007).

In (Chen et al., 2010), higher order Mel-cepstral coefficients (MCEPs) are employed to detect synthetic speech produced by an HMM-based synthesis system. The higher order cepstral coefficients, which reflect the detail in the spectral envelope, tend

to be smoothed during the HMM model parameter training and synthesis processes. Therefore, the higher order cepstral components of synthetic speech exhibit less variance than natural speech. While estimates of this variance thus provide a means of discriminating between genuine and synthetic speech, such an approach is based on the full knowledge of a specific HMM-based speech synthesis system. The same countermeasure may thus not generalise well to other synthesisers which utilise different acoustic parameterisations.

There are some attempts which focus on acoustic differences between vocoders and natural speech. Since the human auditory system is thought to be relatively insensitive to phase (Quatieri, 2002), vocoders typically do not reconstruct speech-like phase information. This simplification leads to differences in the phase spectra between human and synthetic speech, differences which can be utilised for discrimination (De Leon et al., 2012a; Wu et al., 2012a). These methods work well when combined with prior knowledge of the vocoders.

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 (Ogihara et al., 2005; De Leon et al., 2012b). 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.

A summary of ASV performance with integrated spoofing countermeasures is presented in the right-most column of Table 4. While the countermeasure investigated in (De Leon et al., 2012a) is shown to be effective in protecting both GMM-UBM and SVM systems from spoofing, as discussed above, most exploit prior knowledge of specific spoofing algorithms.

5.4. Voice conversion

Voice conversion aims to manipulate the speech of a given speaker so that it resembles in some sense that of another, target speaker (Stylianou, 2009; Evans et al., 2014a). In contrast to speech synthesis systems which require text input, the input to a voice conversion system is a natural speech signal. Typically, voice conversion involves spectral mapping and prosody conversion. Spectral mapping relates to voice timbre, while prosody conversion relates to prosodic features, such as fundamental frequency and duration.

There are three major approaches to spectral mapping: statistical parametric, frequency warping and unit-selection. Statistical parametric approaches usually implement linear or nonlinear conversion functions to map the spectral features of an input speech signal towards features representative of the target speaker. A straightforward approach to spectral mapping based on vector quantisation (VQ) was proposed in (Abe et al., 1988). A mapping codebook is learned from source-target feature pairs and is then used to estimate target features from source features at runtime. Gaussian mixture model (GMM) based approaches, which improve on the hard clustering of VQ methods were proposed in (Kain and Macon, 1998; Stylianou et al., 1998; Toda et al., 2007) to implement a weighted linear conversion function. Alternative, nonlinear approaches include those based on neural networks (Desai et al., 2010), dynamic kernel partial least squares (Helander et al., 2012), restricted Boltzmann machines (Chen et al., 2013) and deep belief networks (Nakashika et al., 2013).

As alternatives to data-driven statistical conversion methods, frequency warping based approaches to voice conversion were introduced in (Toda et al., 2001; Sundermann and Ney, 2003; Erro et al., 2010; Godoy et al., 2012; Erro et al., 2013). Rather than directly substituting the spectral characteristics of the input speech signal, these techniques effectively warp the frequency axis of a source spectrum to match that of the target. Frequency warping approaches tend to retain spectral details and produce high quality converted speech. A so-called Gaussian-dependent filtering approach to voice conversion introduced in (Matrouf et al., 2006; Bonastre et al., 2007) is related to amplitude scaling (Godoy et al., 2012) within a frequency warping framework.

Similar to unit selection speech synthesis, unit selection approaches to voice conversion have also been investigated as a means of directly utilising the target speaker's speech segments to generate converted speech (Sundermann et al., 2006; Dutoit et al., 2007; Wu et al., 2013a). As reported in the voice conversion literature, unit selection approaches produce converted speech much closer to the target speaker than statistical parametric approaches (Sundermann et al., 2006; Dutoit et al., 2007; Wu et al., 2013a) in terms of speaker individuality and subjective listening tests.

In addition to the spectral content, prosody information also plays an important role in characterising speaker individuality. Among the most significant aspects of prosody investigated in the context of voice conversion are the fundamental frequency (F0) and duration. Approaches to convert a source speaker's F0 trajectories to those of a target speaker were investigated in (Gillet and King, 2003; Wu et al., 2006; Helander and Nurminen, 2007; Wu et al., 2010) whereas phoneme or syllable duration conversion approaches were reported in (Wu et al., 2006; Lolive et al., 2008).

Voice conversion technology is likely to be effective in attacking ASV systems. Spectral mapping techniques shift an impostor's spectral characteristics to match those of a specific target speaker and hence present a threat to ASV systems which use spectral features. Meanwhile, prosody conversion can manipulate an attacker's prosodic characteristics to mimic those of a target speaker and thus they present a risk to ASV systems which use prosodic features, e.g. (Adami et al., 2003; Kajarekar et al., 2003; Shriberg et al., 2005; Dehak et al., 2007; Ferrer et al., 2010; Siddiq et al., 2012; Kockmann, 2012).

5.4.1. Spoofing

Voice conversion has attracted increasing interest in the context of ASV spoofing for over a decade. In (Pellom and Hansen, 1999), the vulnerability of a GMM-UBM ASV system was evaluated using the YOHO corpus, which consists of 138 speakers. These experiments showed that the FAR increased from a little over 1 % to 86 % as a result of voice conversion attacks.

Some of the early work in larger-scale, text-independent speaker verification spoofing includes that in (Perrot et al.,

Table 5: Summary of voice conversion spoofing attack and countermeasure (CM) studies. ≈ means the numbers are estimated from the detection error trade-off (DET) curses as presented in the literature.

Study	# target speaker	ASV system	Before spoofing	After spoofing		With CMs
			EER/FAR	EER	FAR	FAR
(Perrot et al., 2005)	n/a	GMM-UBM	≈16 %	26.00 %	≈40 %	n/a
(Matrouf et al., 2006)	n/a	GMM-UBM	≈8 %	≈63 %	≈100 %	n/a
(Bonastre et al., 2007)	n/a	GMM-UBM	6.61 %	28.07 %	≈55 %	n/a
(Kinnunen et al., 2012)	504	JFA	3.24 %	7.61 %	17.33 %	n/a
(Wu et al., 2012b)	504	PLDA	2.99 %	11.18 %	41.25 %	1.71 %
(Alegre et al., 2012b)	201	FA	4.80 %	64.30 %	≈77 %	0 %
(Alegre et al., 2013c)	298	FA	5.60 %	24.40 %	≈54 %	1.60 %
(Alegre et al., 2013a)	298	PLDA	3.03 %	20.2 %	≈55 %	4.10 %
(Kons and Aronowitz, 2013)	750	HMM-NAP	1.00 %	2.90 %	36.00 %	n/a

2005; Matrouf et al., 2006). The work in (Perrot et al., 2005) evaluated the vulnerability of a GMM-UBM ASV system. Experiments reported on the 2004 NIST speaker recognition evaluation (SRE) dataset showed that a baseline EER of 16 % increased to 26 % as a result of voice conversion attacks. The work in (Matrouf et al., 2006) investigated a Gaussian-dependent filtering approach to convert the spectral envelope of the input speech signal towards that of the target speaker. These experiments, conducted on the 2005 NIST SRE dataset, showed that the baseline EER for a GMM-UBM system increased from 8 % to over 60 % as a result of voice conversion attacks which exploit knowledge of the ASV system. The work in (Bonastre et al., 2007), conducted on the 2005 and 2006 NIST SRE datasets, showed a reduced degradation in the EER from 6.61 % to 28.7 % when different feature parameterisations are used for ASV and voice conversion. Even so, this particular approach to voice conversion produces high-quality, natural speech.

The work in (Kinnunen et al., 2012) and (Wu et al., 2012b) extended the study of GMM-UBM systems to consider an array of different approaches to ASV. The work was performed on the 2006 NIST SRE dataset using both joint-density GMM and unit selection approaches to voice conversion. Even if converted speech could be detected easily by human listeners, experiments involving six different ASV systems showed universal susceptibility to spoofing. The FAR of the JFA system increased from 3.24 % to over 17 % in the case of GMM-based voice conversion attacks. That of the most robust PLDA system increased from 2.99 % to over 40 % in the face of unit selection conversion attack. These results are due to the considerable overlap in the distribution of ASV scores for genuine and converted speech, as shown in Figure 9.

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 (Alegre et al., 2012a) and (Alegre et al., 2012b) that certain short intervals of converted speech yield extremely high scores or likelihoods. On their own, such short 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 optimised 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.

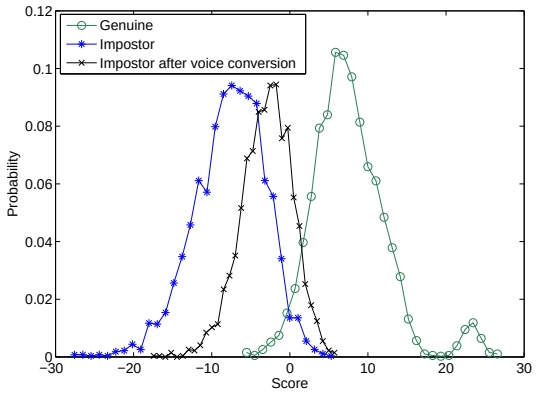


Figure 9: Score distributions before and after voice conversion attacks for a PLDA system as reported in (Wu et al., 2012b).

The work in (Kons and Aronowitz, 2013) 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 hidden Markov model (HMM) instead of a GMM to capture temporal information. 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 1 % to 36 %.

Table 5 presents a summary of spoofing studies described above. Unlike impersonation and replay spoofing studies, and as illustrated in the second column of Table 5, most studies involving voice conversion were performed with large-scale datasets with a large number of speakers. Even though some approaches to voice conversion produce speech with clearly audible artefacts (Chen et al., 2003; Toda et al., 2007; Erro et al., 2013), Table 5 shows that all provoke significant increases in the FAR across a variety of different ASV systems.

5.4.2. Countermeasures

Voice conversion bears some similarity to speech synthesis in that some voice conversion algorithms employ vocoding techniques similar to those used in statistical parametric speech synthesis (Zen et al., 2009). Accordingly, some of the first work

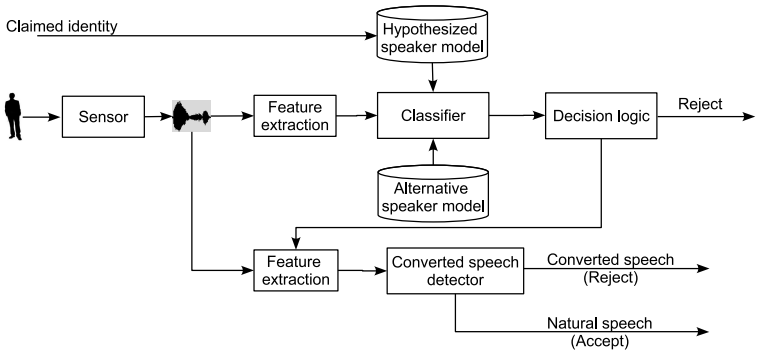


Figure 10: An example of a synthetic speech detector combined with speaker verification (Wu et al., 2012b). Based on prior knowledge that many analysis-synthesis modules used in voice conversion and TTS systems discard natural speech phase, phase characteristics parameterised via modified group delay (MGD) can be used for discriminating natural and synthetic speech.

to detect converted speech drew on related work in synthetic speech detection (De Leon et al., 2011).

The work in (Wu et al., 2012a) exploited artefacts introduced by the vocoder as a means of discriminating converted speech from natural speech. Cosine normalised phase (cos-phase) and modified group delay phase (MGD-phase) features were shown to be effective. Experiments performed on the 2006 NIST SRE dataset were shown to give a detection EER of 5.95 % and 2.35% using cos-phase and MGD-phase countermeasures, respectively. This work was extended in (Wu et al., 2012b) to investigate the effect of countermeasure performance on that of ASV, as illustrated in Figure 10. With the countermeasure, the FAR of a PLDA ASV system reduced from 19.27 % and 41.25% to 0.0 % and 1.71 % under GMM and unit-selection voice conversion spoofing attacks respectively. Interestingly, baseline performance was not affected as a result of integrating spoofing countermeasures. Even so, being based on the absence of natural phase, neither countermeasure is likely to detect converted voice exhibiting real-speech phase, as produced by the conversion approach in (Matrouf et al., 2006).

The work in (Alegre et al., 2012b, 2013b) assessed an approach to detect both voice conversion attacks which preserve real-speech phase (Matrouf et al., 2006; Bonastre et al., 2007) and artificial signal attacks (Alegre et al., 2012a). Results in (Alegre et al., 2012b) suggest that supervector-based SVM classifiers are naturally robust to artificial signal attacks whereas the work in (Alegre et al., 2013b) shows that voice conversion attacks can be detected effectively using estimates of utterance-level, dynamic speech variability. Converted speech was shown to exhibit less dynamic variability than natural speech. The effect of countermeasures on ASV performance was assessed in (Alegre et al., 2013c). The FAR of an FA system was shown to fall from 54 % under spoofing to 2 % with integrated spoofing countermeasures.

A summary of the efforts to develop countermeasures against voice conversion spoofing attacks is presented in Table 5. It shows that countermeasures are effective in protecting ASV systems from voice conversion attacks, and that performance with integrated countermeasures is not too dissimilar to baseline performance.

6. Discussion

As discussed in Section 5, spoofing and countermeasures for ASV have been studied with various approaches to simulate spoofing attacks, different ASV systems, diverse experimental designs, and with a multitude of different datasets, evaluation protocols and metrics. The lack of commonality makes the comparison of vulnerabilities and countermeasure performance extremely challenging. Drawing carefully upon the literature and the authors' own experience, we have nevertheless made such an attempt.

6.1. Spoofing

In Table 6, we summarise the threat of the four major approaches to spoofing considered in this paper. Each attack is compared in terms of *accessibility* and *effectiveness*. *Accessibility* is intended to reflect the ease with which the attack may be performed, i.e. whether the technology is widely known and available or whether it is limited to the technically-knowledgeable. *Effectiveness* reflects the increase in FAR caused by each attack, or the risk it poses to ASV.

Although some studies have shown that impersonation can fool ASV systems, in practice the effectiveness seems to depend on the skill of the impersonator, the similarity of the attacker's voice to that of the target speaker, as well as the system itself. There are clearly easier, more accessible and more effective approaches to spoofing. Indeed, replay attacks are highly effective in the case of text-independent ASV and fixed-phrase text-dependent 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.

Neither speech synthesis nor voice conversion systems capable of producing speech indicative of other specific, target speakers are readily available in commercial off-the-shelf systems. Nonetheless, both speech synthesis and speaker adaptation are active research topics with clear commercial applications. Trainable speech synthesis and publicly available voice conversion tools are already in the public domain, e.g. Festival[6]

[6] `http://www.cstr.ed.ac.uk/projects/festival/`

Table 6: A summary of the accessibility and effectiveness of the four spoofing attack approaches, and the availability of countermeasures for automatic speaker verification. They are graded on a three-level scale.

Spoofing technique	Accessibility (practicality)	Effectiveness (risk)		Countermeasure availability
		Text-independent	Text-dependent	
Impersonation	Low	Low	Low	Non-existent
Replay	High	High	Low to high	Low
Speech synthesis	Medium to high	High	High	Medium
Voice conversion	Medium to high	High	High	Medium

and Festvox[7] and it has been reported that some speech synthesis systems are able to produce speech comparable in quality to human speech[8]. The accessibility of speech synthesis and voice conversion attacks should thus be considered medium to high. Among the others considered in this paper, speech synthesis and voice conversion spoofing attacks may pose the greatest threat to ASV performance and thus effectiveness, for both text-dependent and text-independent ASV systems is high.

6.2. Countermeasures

The vulnerability of ASV systems to each of the four attacks considered above has been confirmed by several independent studies. Even so, efforts to develop countermeasures are relatively embryonic, lagging far behind the level of effort in the case of some other biometric modalities. Also summarised in Table 6 is the current *availability* of countermeasures for each spoofing attack, namely the status of countermeasures for immediate, practical use.

Since impersonated speech is entirely natural, there are no *processing* artefacts which might otherwise be useful for detection purposes. Furthermore, to the best of our knowledge, there are no impersonation countermeasures in the literature and thus the availability is indicated as *non-existent* in Table 6.

Only a small number of countermeasures have been reported in the literature for replay attacks. Availability is thus indicated as low in Table 6. Even if speech synthesis and voice conversion have attracted greater attention, the majority of existing countermeasures make unrealistic use of prior knowledge. Availability is therefore indicated as medium. Furthermore, these countermeasures might be easily overcome if they are known to spoofing attackers. For example, countermeasures based on phase-related features can be overcome by including natural phase information.

6.3. Generalised countermeasures

All of the past work described above targets a specific form of spoofing and generally exploits some prior knowledge of a particular spoofing algorithm. In practice, however, neither the form of spoofing nor the exact algorithm can be known with any certainty. Hence, countermeasures based on processing artefacts indicative of a specific approach to spoofing may not generalise well in the face of varying attacks. Recent work has thus investigated the reliability of specific countermeasures to different attacks (not seen during training) in addition to new, generalised approaches.

The potential for generalised countermeasures is highlighted in independent studies of spoofing with synthetic speech (De Leon et al., 2012a) and converted voice (Wu et al., 2012a). Since both forms of attack employ vocoding techniques, the use of phase information proved a reliable means of detecting manipulated speech signals in both studies. The work in (Wu et al., 2013b) also showed that a common countermeasure based on long-term, temporal magnitude and phase modulation features was successful in detecting both synthetic and converted speech, even if the countermeasure exploits knowledge of the vocoder. Longer-term or higher-level features were investigated in (Alegre et al., 2013c) in the form of local binary pattern (LBP) analysis (Figure 11), a technique originally developed for texture analysis in computer vision problems (Pietikäinen et al., 2011). The LBP-based countermeasure optimised for voice conversion was shown also to be effective in detecting entirely different (no common vocoder) speech synthesis and artificial signal attacks.

While still based on the LBP analysis proposed in (Alegre et al., 2013c), the first entirely generalised countermeasure was proposed in (Alegre et al., 2013a). Generality is ensured through the learning of a one-class classifier optimised using natural speech alone, without any form of spoofed speech training data. Despite the lack of any matched training data, experimental results presented in Figure 12 show that the generalised, one-class classifier is effective in detecting both synthetic and converted speech, in addition to artificial signal spoofing attacks for which the detection EER is 0 %.

7. Issues for future research

As discussed in Section 5, the spoofing and countermeasure studies reported in the literature were conducted with different datasets, evaluation protocols and metrics. Unfortunately, the lack of standards presents a fundamental barrier to the comparison of different research results. This section discusses the current evaluation protocols and metrics and some weaknesses in the methodology. We also discuss some open-source software packages which can be recommended for future spoofing and countermeasure research.

7.1. Large-scale standard datasets

Past studies of impersonation and replay spoofing attacks were all conducted using small-scale datasets, with only small

[7] http://www.festvox.org/index.html

[8] http://www.festvox.org/blizzard/index.html

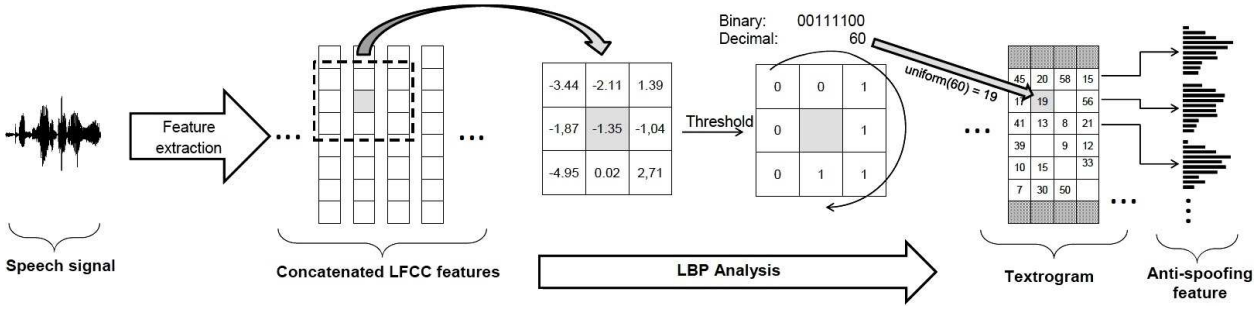


Figure 11: An illustration of local binary pattern (LBP) feature extraction. The procedure involves: a) the extraction of linear frequency cepstral coefficient (LFCC) features from the speech signal; b) the application of uniform LBP analysis to convert the cepstrogram into a so-called textrogram; c) the generation of histograms of LBP values for all but the first and last rows of the textrogram; d) the concatenation of normalised histograms to form feature supervectors for spoofing detection. More details can be found in (Alegre et al., 2013c). Figure reproduced from (Alegre et al., 2013c).

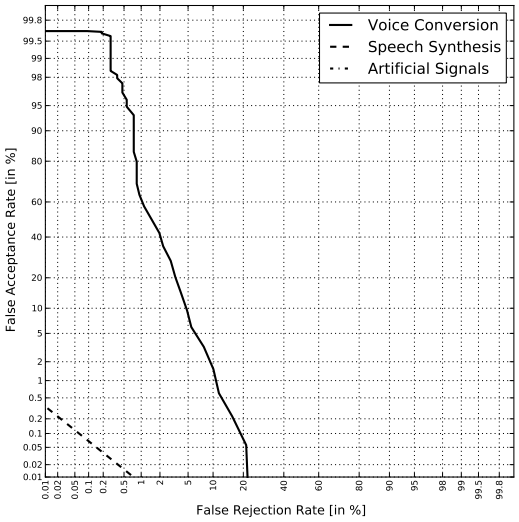


Figure 12: Detection performance for the first generalised, one-class classifier in the case of voice conversion, speech synthesis and artificial signal spoofing attacks. The profile for artificial signals is not visible since the EER is 0 %. Figure reproduced from (Alegre et al., 2013a).

numbers of speakers. While they all suggest that ASV systems might be vulnerable, it is difficult to draw any more meaningful conclusions without the use of significantly larger datasets. While many of the past studies on speech synthesis and voice conversion spoofing attacks already employ large-scale datasets, e.g. NIST speaker recognition evaluation (SRE) corpora, they all require the use of non-standard speech synthesis and voice conversion algorithms in order to generate spoofed speech. Moreover, some of the studies involving speech synthesis have used datasets of high-quality speech recorded in clean, controlled conditions; they lack channel/noise mismatch which might typify the practical use-case scenario. Larger-scale, standard datasets with realistic channel and recording environment variability will be needed for future work in order that the threat from each form of attack can be compared reliably under realistic conditions (Alegre et al., 2014).

It is probably for the study of countermeasures, however, where the need for standard datasets is greatest. All the past work has investigated countermeasures where details of the spoofing attack are either implicitly or explicitly known, e.g. the form of attack or even the exact algorithm. This is clearly wholly unrepresentative of the practical scenario where the nature of the spoofing attack can never be known precisely. In this sense, while past work is sufficient to demonstrate the potential of spoofing countermeasures, their performance is probably over-estimated. In addition, most of the past countermeasure studies have been conducted under matched conditions, e.g. where speech samples used to optimise the countermeasure are collected in the same or similar acoustic environment and over the same or similar channel as those used for evaluation. Large-scale, standard datasets are thus also needed in order that countermeasure performance can be evaluated not only with realistic channel or recording environment variability, but also in the absence of a priori knowledge and hence under variable attacks. The detection of spoofing will then be considerably more challenging but more reflective of practical use cases.

7.2. Evaluation 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 as illustrated in Figure 10. As shown in Table 1, a standard ASV system measures two types of error: false acceptances and false rejections. Similarly, there are also two incorrect outcomes from a stand-alone countermeasure. The assessment of countermeasure performance on its own is relatively straightforward; results are readily analysed with standard detection error trade-off (DET) profiles (Martin et al., 1997) and related metrics.

It is often of interest, however, that the assessment reflects the impact on ASV performance. Assessment is then non-trivial, calling for the joint optimisation of combined classifiers. As reflected in Section 5, there are currently no standard evaluation protocols, metrics or ASV systems and there is thus a need to define such standards in the future.

Candidate standards for evaluation protocols and metrics are being drafted within the scope of the EU FP7 TABULA RASA project[9]. Here, independent countermeasures preceding biometric verification are optimised at three different operating points where thresholds are set to obtain false fake rejection rates (the probability of labelling a genuine access as a spoofing attack) of either 1 %, 5 % or 10 %. Only those samples labelled as genuine accesses are passed to the ASV system whereas those labelled as spoofed accesses are discarded[10]. Performance is assessed using four different DET profiles[11], examples of which are illustrated in Figure 13. The four profiles illustrate performance of the baseline system with naïve (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 impacts of countermeasures. These include those on licit transactions (any deterioration in false rejection – difference between 1st and 2nd profiles) and those on robustness to spoofing (improvements in false acceptance – difference between 3rd and 4th profiles). However, profiles 2 and 4 are dependent on the countermeasure threshold whereas the comparison of profiles 1 and 4 is potentially misleading; they reflect simultaneous changes to both the system and the dataset.

The expected performance and spoofability curve (EPSC[12]) provides an alternative approach to evaluate biometric systems with integrated spoofing countermeasures (Marcel, 2013). The EPSC metric is applied to the fused scores produced by independent biometric and countermeasure classifiers and reflects a trade-off between the half total error rate (HTER) and the so-called spoof false acceptance rate (SFAR). The HTER is the mean of the weighted FAR and FRR (Chingovska et al., 2013) whereas the SFAR refers to the probability of a spoofed access being falsely accepted (Johnson et al., 2010). The HTER is determined with a decision threshold $\tau^\star_\omega$ which minimises the difference between the FRR and the weighted FAR (FAR_ω) according to:

$$\tau^\star_\omega = \arg\min_\tau |\text{FAR}_\omega(\tau, \mathcal{D}_{\text{dev}}) - \text{FRR}(\tau, \mathcal{D}_{\text{dev}})|, \qquad (2)$$

where

$$\text{FAR}_\omega = \omega \cdot \text{SFAR} + (1-\omega) \cdot \text{FAR}, \qquad (3)$$

where ω is the weight which balances the SFAR and FAR, and where $\mathcal{D}_{\text{dev}}$ refers to the development set. Hence, the decision threshold depends on the weight ω.

With the decision threshold $\tau^\star_\omega$, the HTER can be computed on the test set $\mathcal{D}_{\text{test}}$ according to:

$$\text{HTER}_\omega(\tau^\star_\omega) = \frac{\text{FAR}_\omega(\tau^\star_\omega, \mathcal{D}_{\text{test}}) + \text{FRR}(\tau^\star_\omega, \mathcal{D}_{\text{test}})}{2}, \qquad (4)$$

The HTERs are thus computed as a function of ω. The SFAR is similarly computed as a function of ω on the test set. In this way, the EPSC explicitly reflects three types of error metrics, the FAR, FRR and SFAR, while still providing a single combined metric with a unique decision threshold. The EPSC also supports the performance comparison of different countermeasures or ASV systems. However, the EPSC metric is only applicable where the countermeasure and ASV classifiers are fused at the score level. More details of the EPSC can be found in (Marcel, 2013; Chingovska et al., 2014) and an open-source implementation is available in the Bob toolkit (Anjos et al., 2012).

In general, the interpretation of existing evaluation metrics is non-trivial and the metrics themselves lack universal applicability across different approaches to system integration. Further work is thus required to design intuitive, universal metrics which represent the performance of spoofing countermeasures when combined with ASV.

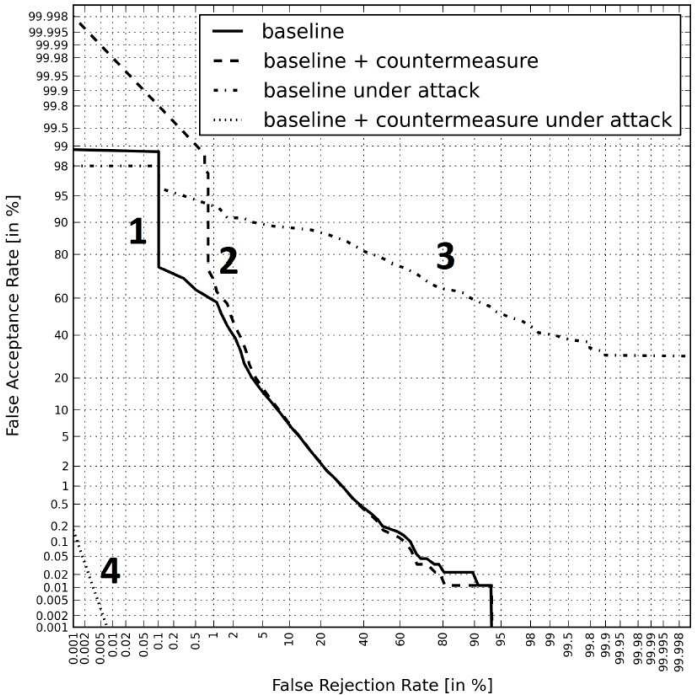


Figure 13: An example of four DET curves needed to analyse vulnerabilities to spoofing and countermeasure performance, both on licit and spoofed access attempts. Results correspond to spoofing attacks using synthetic speech and a standard GMM-UBM classifier assessed on the male subset of the 2006 NIST SRE dataset.

7.3. Open-source software packages

As reflected throughout this article, spoofing and countermeasure studies involve a broad range of technologies, including ASV, speech synthesis and voice conversion. In order to facilitate further research, this section highlights a number of

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

[10] In practice spoofed accesses cannot be fully discarded since so doing would unduly influence ASV false reject and false acceptance rates calculated as a percentage of all accesses. Instead, spoofed accesses bypass ASV and are assigned an arbitrary, low score.

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

[12] `https://pypi.python.org/pypi/antispoofing.evaluation`

useful open-source software packages which can be used either for ASV or spoofing and countermeasure research.

The ALIZE library and associated toolkits[13] are among the most popular in ASV research. Version 3.0 of ALIZE includes several state-of-the-art approaches including joint factor analysis (JFA), i-vector modelling and probabilistic linear discriminant analysis (PLDA) (Larcher et al., 2013a). The Bob signal processing and machine learning toolbox[14], is a general purpose biometric toolkit which also includes ASV functionality (Anjos et al., 2012). Popular solutions for feature extraction include SPro[15] and the Hidden Markov Model Toolkit[16] (HTK) which also includes extensive statistical modelling functionalities.

Some toolkits also provide speech synthesis and voice conversion functionalities. The HMM-based Speech Synthesis System[17] (HTS) can be used to implement HMM-based speech synthesis as well as speaker model adaptation, whereas the Festvox[18] toolkit can be used for voice conversion. The Speech Signal Processing Toolkit[19] (SPTK) offers speech analysis and synthesis functionalities which can be used for feature extraction and the reconstruction of audible speech signals when combined with HTS and Festvox.

7.4. Future directions

The survey highlights the lack of standards which in turn leads to a number of issues in the current methodology, all of which need attention in the future.

Generalised countermeasures: the majority of past anti-spoofing studies have focused on a specific spoofing attack, while variable attacks can be expected in practice. Future research should continue the pursuit of generalised countermeasures capable of detecting different spoofing attacks unseen during countermeasure optimisation. Such work may potentially build on the one-class approach (Alegre et al., 2013a) where the countermeasure is trained only with natural speech. Evaluation protocols should include diverse, mismatched spoofing techniques thereby reflecting the uncertainty in the likely nature of a spoofing attack.

Text-dependent systems: on account of dataset availability, the majority of past work involves text-independent ASV which is arguably more relevant to surveillance applications. Future work should increase the focus on text-dependent systems, more pertinent to authentication scenarios.

Replay attacks: the present literature focuses on relatively sophisticated attacks, such as synthesised and converted speech. With the expertise to implement such attacks being beyond the means of the lay person, greater emphasis should be placed on the less effective, though more accessible attacks; even if they are less effective, they might occur more frequently in practice. The most obvious, accessible attack involves replay.

Countermeasures under acoustic mismatch: most evaluations reported to date involve speech data with channel and recording environment variability identical to that used in countermeasure optimisation. Different transmission channels, additive noises and other imperfections should be expected in practice and have potential to mask processing artefacts key to spoofing detection. Future evaluations should thus evaluate countermeasures under acoustically degraded and channel-mismatched conditions.

Combined spoofing attacks: The majority of the past studies involve only independent spoofing approaches. Future work should consider the possibility of attackers combining several spoofing techniques to boost effectiveness. For example, voice conversion and impersonation could be combined to spoof both spectral and prosodic cues.

It is, however, the consistent theme throughout this article, namely the lack of standard databases, protocols and metrics, which leads to what is arguably the most urgent of all directions for the future. The use of different experimental configurations impedes the comparison of different results and will be a fundamental barrier to future advances; such standards are essential to the benchmarking of different ideas and experience shows they are critical to progress. Just as they have been for progress in automatic speaker verification, standard databases, protocols and metrics will be an essential component in the future for spoofing and countermeasure research. First and foremost, the future work should define a publicly available dataset and competitive challenge similar in spirit to the traditional NIST speaker recognition evaluations. The authors are currently working in this direction.

8. Conclusions

This article reviews the previous work to assess the vulnerability of automatic speaker verification systems to spoofing and the potential to protect them using dedicated countermeasures. Even if there are currently no standard datasets, evaluation protocols or metrics with which to conduct meaningfully comparable or reproducible research, previous studies involving impersonation, replay, speech synthesis and voice conversion all indicate genuine vulnerabilities. While a growing body of independent research also points to the potential of countermeasures, fundamental shortcomings in the research methodology are common to all the past work and point towards specific priorities for the future. Finally, while there is potential for next generation countermeasures to detect varying spoofing attacks, a continuous arms race is likely; efforts to develop more sophisticated countermeasures will likely be accompanied by increased efforts to spoof automatic speaker verification systems.

[13] http://alize.univ-avignon.fr/
[14] http://idiap.github.io/bob/
[15] http://www.irisa.fr/metiss/guig/spro/
[16] http://htk.eng.cam.ac.uk/
[17] http://hts.sp.nitech.ac.jp/
[18] http://festvox.org/
[19] http://sp-tk.sourceforge.net/

The area is therefore set to remain an important field of research in the future.

Acknowledgements

This work was partially supported by Academy of Finland (project number 253120, 283256) and by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU) (grant agreement number 257289). The third author would like to thank Prof. Anne-Maria Laukkanen and M.Sc. Johanna Leskelä at the University of Tampere for the permission to use impersonation data.

References

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

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

Akhtar, Z., Fumera, G., Marcialis, G.L., Roli, F., 2012. Evaluation of serial and parallel multibiometric systems under spoong attacks, in: Proc. 5th Int. Conf. on Biometrics (ICB 2012).

Alegre, F., Amehraye, A., Evans, N., 2013a. 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).

Alegre, F., Amehraye, A., Evans, N., 2013b. Spoofing countermeasures to protect automatic speaker verification from voice conversion, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Alegre, F., Evans, N., Kinnunen, T., Wu, Z., Yamagishi, J., 2014. Anti-spoofing: voice databases, in: Li, S.Z., Jain, A.K. (Eds.), Encyclopedia of biometrics. Springer-Verlag, US.

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

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

Alegre, F., Vipperla, R., Evans, N., et al., 2012b. Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals, in: Proc. Interspeech.

Amin, T.B., German, J.S., Marziliano, P., 2013. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. Journal of the Acoustic Society of America 134, 4068–4068.

Amin, T.B., Marziliano, P., German, J.S., 2014. Glottal and vocal tract characteristics of voice impersonators. IEEE Trans. on Multimedia 16, 668–678.

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

Beutnagel, B., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A., 1999. The AT&T Next-Gen TTS system, in: Proc. Joint ASA, EAA and DAEA Meeting.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A., 2004. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing 2004, 430–451.

Black, A.W., 2006. CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling, in: Proc. Interspeech.

Blomberg, M., Elenius, D., Zetterholm, E., 2004. Speaker verification scores and acoustic analysis of a professional impersonator, in: Proc. FONETIK.

Boersma, P., Weenink, D., 2014. Praat: doing phonetics by computer. Computer program. Version 5.3.64, retrieved 12 February 2014 from `http://www.praat.org/`.

Bonastre, J.F., Matrouf, D., Fredouille, C., 2007. Artificial impostor voice transformation effects on false acceptance rates, in: Proc. Interspeech.

Bredin, H., Miguel, A., Witten, I.H., Chollet, G., 2006. Detecting replay attacks in audiovisual identity verification, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Breen, A., Jackson, P., 1998. A phonologically motivated method of selecting nonuniform units, in: Proc. Int. Conf. on Spoken Language Processing (ICSLP).

Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., Leeuwen, D., Matějka, P., Schwartz, P., Strasheim, A., 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans. Audio, Speech and Language Processing 15, 2072–2084.

Burget, L., Matějka, P., Schwarz, P., Glembek, O., Černocký, J., 2007. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans. Audio, Speech and Language Processing 15, 1979–1986.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13, 308–311.

Campbell Jr, J.P., 1997. Speaker recognition: A tutorial. Proceedings of the IEEE 85, 1437–1462.

Chen, L.H., Ling, Z.H., Song, Y., Dai, L.R., 2013. Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion, in: Proc. Interspeech.

Chen, L.W., Guo, W., Dai, L.R., 2010. Speaker verification against synthetic speech, in: 7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP).

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

Chingovska, I., Anjos, A., Marcel, S., 2013. Anti-spoofing in action: joint operation with a verification system, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW).

Chingovska, I., Anjos, A., Marcel, S., 2014. Biometrics evaluation under spoofing attacks. IEEE Trans. on Information Forensics and Security .

Coorman, G., Fackrell, J., Rutten, P., Coile, B., 2000. Segment selection in the L & H realspeak laboratory TTS system, in: Proc. Int. Conf. on Spoken Language Processing (ICSLP), pp. 395–398.

De Leon, P.L., Apsingekar, V.R., Pucher, M., Yamagishi, J., 2010a. Revisiting the security of speaker verification systems against imposture using synthetic speech, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

De Leon, P.L., Hernaez, I., Saratxaga, I., Pucher, M., Yamagishi, J., 2011. Detection of synthetic speech for the problem of imposture, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

De Leon, P.L., Pucher, M., Yamagishi, J., 2010b. Evaluation of the vulnerability of speaker verification to synthetic speech, in: Proc. Odyssey: the Speaker and Language Recognition Workshop.

De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I., 2012a. Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans. Audio, Speech and Language Processing 20, 2280–2290.

De Leon, P.L., Stewart, B., Yamagishi, J., 2012b. Synthetic speech discrimination using pitch pattern statistics derived from image analysis, in: Proc. Interspeech.

Dehak, N., Dumouchel, P., Kenny, P., 2007. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio, Speech and Language Processing 15, 2095–2103.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech and Language Processing 19, 788–798.

Desai, S., Black, A., Yegnanarayana, B., Prahallad, K., 2010. Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio, Speech and Language Processing 18, 954–964.

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).

Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D., 1998. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation, Gaithersburg, MD. National Institute of Standards and Technology.

Donovan, R.E., Eide, E.M., 1998. The IBM trainable speech synthesis system,

in: Proc. Int. Conf. on Spoken Language Processing (ICSLP).
Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., Stylianou, Y., 2007. Towards a voice conversion system based on frame selection, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).
Endres, W., Bambach, W., Flösser, G., 1971. Voice spectrograms as a function of age, voice disguise, and voice imitation. Journal of the Acoustic Society of America 49, 1842–1848.
Eriksson, A., Wretling, P., 1997. How flexible is the human voice? - a case study of mimicry, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).
Erro, D., Moreno, A., Bonafonte, A., 2010. Voice conversion based on weighted frequency warping. IEEE Trans. Audio, Speech and Language Processing 18, 922–931.
Erro, D., Navas, E., Hernaez, I., 2013. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. IEEE Trans. Audio, Speech and Language Processing 21, 556–566.
Evans, N., Alegre, F., Wu, Z., Kinnunen, T., 2014a. Anti-spoofing: voice conversion, in: Li, S.Z., Jain, A.K. (Eds.), Encyclopedia of biometrics. Springer-Verlag, US.
Evans, N., Kinnunen, T., Yamagishi, J., 2013. Spoofing and countermeasures for automatic speaker verification, in: Proc. Interspeech.
Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., DeLeon, P., 2014b. Speaker recognition anti-spoofing, in: Marcel, S., Li, S.Z., Nixon, M. (Eds.), Handbook of biometric anti-spoofing. Springer.
Farrús, M., Wagner, M., Anguita, J., Hernando, J., 2008. How vulnerable are prosodic features to professional imitators?, in: Proc. Odyssey: the Speaker and Language Recognition Workshop.
Farrús, M., Wagner, M., Erro, D., Hernando, J., 2010. Automatic speaker recognition as a measurement of voice imitation and conversion. International Journal of Speech Language and the Law 17, 119–142.
Faundez-Zanuy, M., 2004. On the vulnerability of biometric security systems. IEEE Aerospace and Electronic Systems Magazine 19, 3–8.
Faundez-Zanuy, M., Hagmüller, M., Kubin, G., 2006. Speaker verification security improvement by means of speech watermarking. Speech Communication 48, 1608–1619.
Ferrer, L., Scheffer, N., Shriberg, E., 2010. A comparison of approaches for modeling prosodic features in speaker recognition, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).
Foomany, F., Hirschfield, A., Ingleby, M., 2009. Toward a dynamic framework for security evaluation of voice verification systems, in: Proc. IEEE Toronto Int. Conf. Science and Technology for Humanity (TIC-STH), pp. 22–27.
Galbally, J., McCool, C., Fierrez, J., Marcel, S., Ortega-Garcia, J., 2010. On the vulnerability of face verification systems to hill-climbing attacks. Pattern Recognition 43, 1027–1038.
Galou, G., 2011. Synthetic voice forgery in the forensic context: a short tutorial, in: Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG), pp. 1–3.
Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems, in: Proc. Interspeech.
Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. on Speech and Audio Processing 2, 291–298.
Gerhard, D., 2003. Pitch extraction and fundamental frequency: History and current techniques. Technical Report TR-CS 2003-06. Department of Computer Science, University of Regina.
Gillet, B., King, S., 2003. Transforming F0 contours, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).
Godoy, E., Rosec, O., Chonavel, T., 2012. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. IEEE Trans. Audio, Speech and Language Processing 20, 1313–1323.
Hatch, A.O., Kajarekar, S., Stolcke, A., 2006. Within-class covariance normalization for SVM-based speaker recognition, in: Proc. Int. Conf. on Spoken Language Processing (ICSLP).
Hautamäki, R.G., Kinnunen, T., Hautamäki, V., Laukkanen, A.M., 2014. Comparison of human listeners and speaker verification systems using voice mimicry data, in: Proc. Odyssey: the Speaker and Language Recognition Workshop, Joensuu, Finland. pp. 137–144.
Hautamäki, R.G., Kinnunen, T., Hautamäki, V., Leino, T., Laukkanen, A.M., 2013a. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry, in: Proc. Interspeech.
Hautamäki, V., Kinnunen, T., Sedlák, F., Lee, K.A., Ma, B., Li, H., 2013b. Sparse classifier fusion for speaker verification. IEEE Trans. Audio, Speech and Language Processing 21, 1622–1631.
Helander, E., Silén, H., Virtanen, T., Gabbouj, M., 2012. Voice conversion using dynamic kernel partial least squares regression. IEEE Trans. Audio, Speech and Language Processing 20, 806–817.
Helander, E.E., Nurminen, J., 2007. A novel method for prosody prediction in voice conversion, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29, 82–97.
Hunt, A., Black, A.W., 1996. 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).
Jain, A.K., Ross, A., Pankanti, S., 2006. Biometrics: a tool for information security. IEEE Trans. on Information Forensics and Security 1, 125–143.
Johnson, P., Tan, B., Schuckers, S., 2010. Multimodal fusion vulnerability to non-zero effort (spoof) imposters, in: IEEE Int. Workshop on Information Forensics and Security (WIFS).
Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).
Kajarekar, G.S., Stolcke, E.S.K.S.A., Venkataraman, A., 2003. Modeling duration patterns for speaker recognition, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).
Kenny, P., 2006. Joint factor analysis of speaker and session variability: theory and algorithms. technical report CRIM-06/08-14.
Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio, Speech and Language Processing 15, 1448–1460.
Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of inter-speaker variability in speaker verification. IEEE Trans. Audio, Speech and Language Processing 16, 980–988.
Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52, 12–40.
Kinnunen, T., Wu, Z.Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. 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).
Kitamura, T., 2008. Acoustic analysis of imitated voice produced by a professional impersonator, in: Proc. Interspeech.
Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. Journal of the Acoustic Society of America 67, 971–995.
Kockmann, M., 2012. Subspace Modeling of Prosodic Features for Speaker Verification. Ph.D. thesis. BRNO UNIVERSITY OF TECHNOLOGY. Brno, Czech Republic.
Kons, Z., Aronowitz, H., 2013. Voice transformation-based spoofing of text-dependent speaker verification systems, in: Proc. Interspeech.
Larcher, A., Bonastre, J.F., Fauve, B., Lee, K.A., Lévy, C., Li, H., Mason, J.S., Parfait, J.Y., ValidSoft Ltd, U., 2013a. Alize 3.0-open source toolkit for state-of-the-art speaker recognition, in: Proc. Interspeech.
Larcher, A., Lee, K.A., Ma, B., Li, H., 2012. RSR2015: Database for text-dependent speaker verification using multiple pass-phrases., in: Proc. Interspeech.
Larcher, A., Lee, K.A., Ma, B., Li, H., 2013b. Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).
Larcher, A., Lee, K.A., Ma, B., Li, H., 2014. Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication 60, 5677.
Lau, Y., Tran, D., Wagner, M., 2005. Testing voice mimicry with the YOHO speaker verification corpus, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer. pp. 907–907.
Lau, Y.W., Wagner, M., Tran, D., 2004. Vulnerability of speaker verification to voice mimicking, in: Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Processing.
Lee, K.A., Ma, B., Li, H., 2013. Speaker verification makes its debut in smartphone, in: IEEE Signal Processing Society Speech and language Technical Committee Newsletter,.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171–185.

Leskelä, J., 2011. Changes in F0, formant frequencies and spectral slope in imitation. Master's thesis. University of Tampere. Finland. In Finnish.

Li, H., Ma, B., 2010. Techware: Speaker and spoken language recognition resources [best of the web]. IEEE Signal Processing Magazine 27, 139–142.

Li, H., Ma, B., Lee, K.A., 2013. Spoken language recognition: From fundamentals to practice. Proceedings of the IEEE 101, 1136–1159.

Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J., 2012. Probabilistic models for inference about identity. IEEE Trans. on Pattern Analysis and Machine Intelligence 34, 144–157.

Lindberg, J., Blomberg, M., et al., 1999. Vulnerability in speaker verification-a study of technical impostor techniques, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).

Ling, Z.H., Deng, L., Yu, D., 2013. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. IEEE Trans. Audio, Speech and Language Processing 21, 2129–2139.

Ling, Z.H., Wu, Y.J., Wang, Y.P., Qin, L., Wang, R.H., 2006. USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method, in: the Blizzard Challenge Workshop.

Ling, Z.H., Xia, X.J., Song, Y., Yang, C.Y., Chen, L.H., Dai, L.R., 2012. The USTC system for Blizzard Challenge 2012, in: Blizzard Challenge workshop.

Lolive, D., Barbot, N., Boeffard, O., 2008. Pitch and duration transformation with non-parallel data, in: Proc. Speech Prosody.

Lu, H., King, S., Watts, O., 2013. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis, in: Proc. the 8th ISCA Speech Synthesis Workshop.

Marcel, S., 2013. Spoofing and anti-spoofing in biometrics: Lessons learned from the tabula rasa project. Tutorial. Retrieved 26 February 2014 from `http://www.idiap.ch/~marcel/professional/BTAS_2013.html`.

Mariéthoz, J., Bengio, S., 2006. Can a professional imitator fool a GMM-based speaker verification system? IDIAP Research Report (No. Idiap-RR 05-61).

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).

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

Masuko, T., Tokuda, K., Kobayashi, T., 2000. Imposture using synthetic speech against speaker verification based on spectrum and pitch, in: Proc. Interspeech.

Masuko, T., Tokuda, K., Kobayashi, T., Imai, S., 1996. Speech synthesis using HMMs with dynamic features, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Masuko, T., Tokuda, K., Kobayashi, T., Imai, S., 1997. Voice characteristics conversion for HMM-based speech synthesis system, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

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

Matsui, T., Furui, S., 1995. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. Speech Communication 17, 109–116.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9, 453–467.

Nakashika, T., Takashima, R., Takiguchi, T., Ariki, Y., 2013. Voice conversion in high-order eigen space using deep belief nets, in: Proc. Interspeech.

Nuance, 2013. Nuance vocalpassword, in: `http://www.nuance.com/landing-pages/products/voicebiometrics/vocalpassword.asp`.

Ogihara, A., Unno, H., Shiozakai, A., 2005. Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification. IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences 88, 280–286.

Panjwani, S., Prakash, A., 2014. Finding impostors in the crowd: The use of crowdsourcing to attack biometric systems. Unpublished manuscript, Bell Labs India.

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

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

Pietikäinen, M., Hadid, A., Zhao, G., 2011. Computer Vision Using Local Binary Patterns. Springer.

Prince, S., Elder, J., 2007. Probabilistic linear discriminant analysis for inferences about identity, in: Proc. IEEE Int. Conf. on Computer Vision (ICCV).

Qian, Y., Fan, Y., Hu, W., Soong, F.K., 2014. On the training aspects of deep neural network (dnn) for parametric tts synthesis, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Qian, Y., Wu, Z., Gao, B., Soong, F.K., 2011. Improved prosody generation by maximizing joint probability of state and longer units. IEEE Trans. Audio, Speech and Language Processing 19, 1702–1710.

Quatieri, T.F., 2002. Discrete-Time Speech Signal Processing Principles and Practice. Prentice-Hall, Inc.

Ratha, N.K., Connell, J.H., Bolle, R.M., 2001. Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal 40, 614–634.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B., 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41.

Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing 3, 72–83.

Riera, A., Soria-Frisch, A., Acedo, J., Hadid, A., Alegre, F., Evans, N., Marcialis, G.L., 2012. Evaluation of initial non-ICAO countermeasures for spoofing attacks). Technical Report Deliverable D4.2. Trusted biometrics under spoofing attacks (TABULA RASA), 7th Framework Programme of the European, grant agreement number 257289.

Rodriques, R.N., Ling, L.L., Govindaraju, V., 2009. Robustness of multimodal biometric fusion methods agaisnt spoof attacks. Journal of Visual Languages and Computing 20, 169–179.

Saeidi, R., et al., 2013. I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification, in: Proc. Interspeech.

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

Shang, W., Stevenson, M., 2010. Score normalization in playback attack detection, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., 2005. Modeling prosodic feature sequences for speaker recognition. Speech Communication 46, 455–472.

Siddiq, S., Kinnunen, T., Vainio, M., Werner, S., 2012. Intonational speaker verification: a study on parameters and performance under noisy conditions, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

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

Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., Dumouchel, P., 2013. Text-dependent speaker recognition using PLDA with uncertainty propagation, in: Proc. Interspeech.

Stoll, L., Doddington, G., 2010. Hunting for wolves in speaker recognition, in: Proc. Odyssey: the Speaker and Language Recognition Workshop.

Stylianou, Y., 2009. Voice transformation: a survey, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Stylianou, Y., Cappé, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. on Speech and Audio Processing 6, 131–142.

Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A., Narayanan, S., 2006. Text-independent voice conversion based on unit selection, in: Proc.

IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Sundermann, D., Ney, H., 2003. VTLN-based voice conversion, in: Proc. the 3rd IEEE Int. Symposium on Signal Processing and Information Technology.

Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech and Language Processing 15, 2222–2235.

Toda, T., Saruwatari, H., Shikano, K., 2001. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Togneri, R., Pullella, D., 2011. An overview of speaker identification: accuracy and robustness issues. IEEE Circuits and Systems Magazine 11, 23–61.

Tomoki, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. on Information and Systems 90, 816–824.

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

Villalba, J., Lleida, E., 2011a. Detecting replay attacks from far-field recordings on speaker verification systems, in: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N., Fairhurst, M. (Eds.), Biometrics and ID Management. Springer. Lecture Notes in Computer Science, pp. 274–285.

Villalba, J., Lleida, E., 2011b. Preventing replay attacks on speaker verification systems, in: IEEE Int. Carnahan Conf. on Security Technology (ICCST).

Wang, Z.F., Wei, G., He, Q.H., 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition, in: Proc. IEEE Int. Conf. Machine Learning and Cybernetics (ICMLC).

Woodland, P.C., 2001. Speaker adaptation for continuous density HMMs: A review, in: Proc. ISCA Workshop on Adaptation Methods for Speech Recognition.

Wu, C.H., Hsia, C.C., Liu, T.H., Wang, J.F., 2006. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Trans. Audio, Speech and Language Processing 14, 1109–1116.

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

Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E., 2012b. 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 Conference (APSIPA ASC).

Wu, Z., Li, H., 2013. Voice conversion and spoofing attack on speaker verification systems, in: Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC).

Wu, Z., Virtanen, T., Kinnunen, T., Chng, E.S., Li, H., 2013a. Exemplar-based unit selection for voice conversion utilizing temporal information, in: Proc. Interspeech.

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

Wu, Z.Z., Kinnunen, T., Chng, E.S., Li, H., 2010. Text-independent F0 transformation with non-parallel data for voice conversion, in: Proc. Interspeech.

Yager, N., Dunstone, T., 2010. The biometric menagerie. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 220–230.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio, Speech and Language Processing 17, 66–83.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in: Proc. European Conference on Speech Communication and Technology (Eurospeech).

Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks, in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Zen, H., Toda, T., Nakamura, M., Tokuda, K., 2007. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. on Information and Systems E90-D, 325–333.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Communication 51, 1039–1064.

Zetterholm, E., Blomberg, M., Elenius, D., 2004. A comparison between human perception and a speaker verification system score of a voice imitation, in: Proc. of Tenth Australian Int. Conf. on Speech Science & Technology.