# Introduction to Voice Presentation Attack Detection and Recent Advances

Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi and Kong-Aik Lee

**Abstract** Over the past few years significant progress has been made in the field of presentation attack detection (PAD) for automatic speaker recognition (ASV). This includes the development of new speech corpora, standard evaluation protocols and advancements in front-end feature extraction and back-end classifiers. The use of standard databases and evaluation protocols has enabled for the first time the meaningful benchmarking of different PAD solutions. This chapter summarises the progress, with a focus on studies completed in the last three years. The article presents a summary of findings and lessons learned from two ASVspoof challenges, the first community-led benchmarking efforts. These show that ASV PAD remains an unsolved problem and that further attention is required to develop generalised

Md Sahidullah
School of Computing, University of Eastern Finland (Finland), e-mail: sahid@cs.uef.fi
[Currently with Inria, France.]

Héctor Delgado
Department of Digital Security, EURECOM (France) e-mail: hector.delgado@eurecom.fr

Massimiliano Todisco
Department of Digital Security, EURECOM (France) e-mail: massimiliano.todisco@eurecom.fr

Tomi Kinnunen
School of Computing, University of Eastern Finland (Finland), e-mail: tkinnu@cs.uef.fi

Nicholas Evans
Department of Digital Security, EURECOM (France) e-mail: evans@eurecom.fr

Junichi Yamagishi
National Institute of Informatics (Japan) and University of Edinburgh (United Kingdom) e-mail: jyamagis@nii.ac.jp

Kong-Aik Lee
Data Science Research Laboratories, NEC Corporation (Japan) e-mail: k-lee@ax.jp.nec.com

PAD solutions which have potential to detect diverse and previously unseen spoofing attacks.

## 1 Introduction

Automatic speaker verification (ASV) technology aims to recognise individuals using samples of the human voice signal [1, 2]. Most ASV systems operate on estimates of the spectral characteristics of voice in order to recognise individual speakers. ASV technology has matured in recent years and now finds application in a growing variety of real-world authentication scenarios involving both *logical* and *physical* access. In scenarios, ASV technology can be used for remote person authentication via the Internet or traditional telephony. In many cases, ASV serves as a convenient and efficient alternative to more conventional password-based solutions, one prevalent example being person authentication for Internet and mobile banking. scenarios include the use of ASV to protect personal or secure/sensitive facilities, such as domestic and office environments. With the growing, widespread adoption of smartphones and voice-enabled smart devices, such as intelligent personal assistants all equipped with at least one microphone, ASV technology stands to become even more ubiquitous in the future.

Despite its appeal, the now-well-recognised vulnerability to manipulation through presentation attacks (PAs), also known as spoofing, has dented confidence in ASV technology. As identified in ISO/IEC 30107-1 standard [3], the possible locations of presentation attack points in a typical ASV system are illustrated in Fig. 1. Two of the most vulnerable places in an ASV system are marked by 1 and 2, corresponding to physical access and logical access. This work is related to these two types of attacks.

Unfortunately, ASV is arguably more prone to PAs than other biometric systems based on traits or characteristics that are less-easily acquired; samples of a given person's voice can be collected readily by fraudsters through face-to-face or telephone conversations and then replayed in order to manipulate an ASV system. Replay attacks are furthermore only one example of ASV PAs. More advanced voice conversion or speech synthesis algorithms can be used to generate particularly effective PAs using only modest amounts of voice data collected from a target person.

There are a number of ways to prevent PA problems. The first one is based on a text-prompted system which uses an utterance verification process [4]. The user needs to utter a specific text, prompted for authentication by the system which requires a text-verification system. Secondly, as human can never reproduce an identical speech signal, some countermeasures use template matching or audio fingerprinting to verify whether the speech utterance was presented to the system earlier [5]. Thirdly, some work looks into statistical acoustic characterisation of authentic speech and speech created with presentation attack methods or spoofing techniques [6]. Our focus is on the last category, which is more convenient in a practical scenario for both text-dependent and text-independent ASV. In this case,

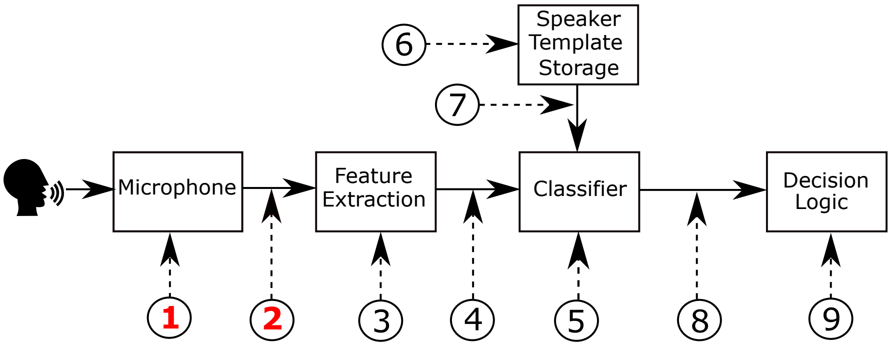

Fig. 1: Possible attack locations in a typical ASV system. 1: microphone point, 2: transmission point, 3: override feature extractor, 4: modify probe to features, 5: override classifier, 6: modify speaker database, 7: modify biometric reference, 8: modify score and 9: override decision.

given a speech signal, $S$, PA detection  here, the determination of whether $S$ is a natural or PA speech  can be formulated as a hypothesis test:

- $H_0$: $S$ is natural speech.
- $H_1$: $S$ is created with PA methods.

A can be applied to decide between $H_0$ and $H_1$. Suppose that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ are the acoustic feature vectors of $N$ speech frames extracted from $S$, then the logarithmic likelihood ratio score is given by,

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_{H_0}) - \log p(\mathbf{X}|\lambda_{H_1}) \quad (1)$$

In1, $\lambda_{H_0}$ and $\lambda_{H_1}$ are the acoustic models to characterise the hypotheses correspondingly for natural speech and PA speech. The parameters of these models are estimated using training data for natural and PA speech. A typical PAD system is shown in Fig. 2. A test speech can be accepted as natural or rejected as PA speech with help of a threshold, $\theta$ computed on some development data. If the score is greater than or equal to the threshold, it is accepted; otherwise, rejected. The performance of the PA system is assessed by computing the (EER) metric. This is the error rate for a specific value of a threshold where two error rates, i.e., the probability of a PA speech detected as being natural speech (known as false acceptance rate or FAR) and the probability of a natural speech speech being misclassified as a PA speech (known as false rejection rate or FRR), are equal. Sometimes (HTER) is also computed [7]. This is the average of FAR and FRR which are computed using a decision threshold obtained with the help of the development data.

Awareness and acceptance of the vulnerability to PAs have generated a growing interest in develop solutions to presentation attack detection (PAD), also referred to as spoofing countermeasures. These are typically dedicated auxiliary systems which function in tandem to ASV in order to detect and deflect PAs. The research in this direction has progressed rapidly in the last three years, due partly to the release of

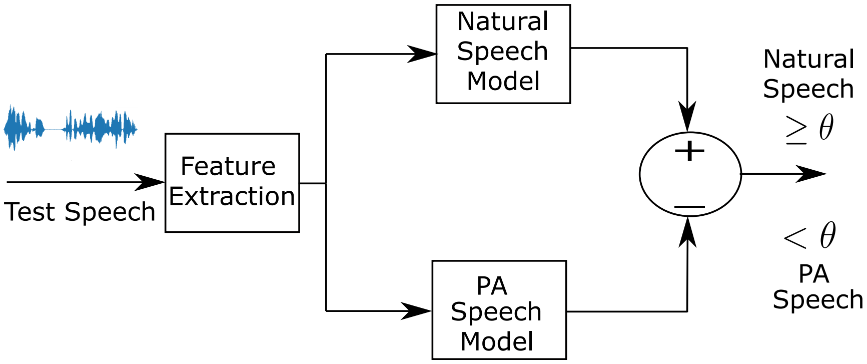

Fig. 2: Block diagram of a typical presentation attack detection system.

several public speech corpora and the organisation of PAD challenges for ASV. This article, a continuation of the chapter [8] in the first edition of the Handbook for Biometrics [9] presents an up-to-date review of the different forms of voice presentation attacks, broadly classified in terms of impersonation, replay, speech synthesis and voice conversion. The primary focus is nonetheless on the progress in PAD. The chapter reviews the most recent work involving a variety of different features and classifiers. Most of the work covered in the chapter relates to that conducted using the two most popular and publicly available databases, which were used for the two ASVspoof challenges co-organized by the authors. The chapter concludes with a discussion of research challenges and future directions in PAD for ASV.

## 2 Basics of ASV spoofing and countermeasures

Spoofing or presentation attacks are performed on a biometric system at the sensor or acquisition level to bias score distributions toward those of genuine clients, thus provoking increases in the false acceptance rate (FAR). This section reviews four well-known ASV spoofing techniques and their respective countermeasures: impersonation, replay, speech synthesis and voice conversion. Here, we mostly review the work in the pre-ASVspoof period, as well as some very recent studies on presentation attacks.

### *2.1 Impersonation*

In speech or mimicry attacks, an intruder speaker intentionally modifies his or her speech to sound like the target speaker. Impersonators are likely to copy lexical, prosodic, and idiosyncratic behaviour of their target speakers presenting a potential point of vulnerability concerning speaker recognition systems.

### 2.1.1 Spoofing

There are several studies about the consequences of mimicry on ASV. Some studies concern attention to the voice modifications performed by professional impersonators. It has been reported that impersonators are often particularly able to adapt the fundamental frequency (F0) and occasionally also the formant frequencies towards those of the target speakers [10, 11, 12]. In studies, the focus has been on analysing the vulnerability of speaker verification systems in the presence of voice mimicry. The studies by Lau et al. [13, 14] suggest that if the target of impersonation is known in advance and his or her voice is “similar” to the impersonator’s voice (in the sense of automatic speaker recognition score), then the chance of spoofing an automatic recognizer is increased. In [15], the experiments indicated that professional impersonators are potentially better impostors than amateur or naive ones. Nevertheless, the voice impersonation was not able to spoof the ASV system. In [10], the authors attempted to quantify how much a speaker is able to approximate other speakers’ voices by selecting a set of prosodic and voice source features. Their prosodic and acoustic based ASV results showed that two professional impersonators imitating known politicians increased the identification error rates.

More recently, a fundamentally different study was carried out by Panjwani et al. [16] using crowdsourcing to recruit both amateur and more professional impersonators. The results showed that impersonators succeed in increasing their average score, but not in exceeding the target speaker score. All of the above studies analysed the effects of speech impersonation either at the acoustic or speaker recognition score level, but none proposed any countermeasures against impersonation. In a recent study [17], the experiments aimed to evaluate the vulnerability of three modern speaker verification systems against impersonation attacks and to further compare these results to the performance of non-expert human listeners. It is observed that, on average, the mimicry attacks lead to increased error rates. The increase in error rates depends on the impersonator and the ASV system.

The main challenge, however, is that no large speech corpora of impersonated speech exists for the quantitative study of impersonation effects on the same scale as for other attacks, such as text-to-speech synthesis and voice conversion, where generation of simulated spoofing attacks as well as developing appropriate countermeasures is more convenient.

### 2.1.2 Countermeasures

While the threat of impersonation is not fully understood due to limited studies involving small datasets, it is perhaps not surprising that there is no prior work investigating 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.

## *2.2 Replay*

attacks refer to the use of pre-recorded speech from a target speaker, which is then replayed through some playback device to feed the system microphone. These attacks require no specific expertise nor sophisticated equipment, thus they are easy to implement. Replay is a relatively low-technology attack within the grasp of any potential attacker even without specialised knowledge in speech processing. Several works in the earlier literature report significant increases in error rates when using replayed speech. Even if replay attacks may present a genuine risk to ASV systems, the use of prompted-phrase has the potential to mitigate the impact.

### 2.2.1 Spoofing

The study on the impact of replay attack on ASV performance was very limited until recently before the release of AVspoof [18] and ASVspoof 2017 corpus. The earlier studies were conducted either on simulated or on real replay recording from far-field.

The vulnerability of ASV systems to replay attacks was first investigated in a text-dependent scenario [19], where the concatenation of recorded digits was tested against a hidden Markov model (HMM) based ASV system. Results showed an increase in the FAR from 1 to 89% for male speakers and from 5 to 100% for female speakers.

The work in [20] investigated text-independent ASV vulnerabilities through the replaying of far-field recorded speech in a mobile telephony scenario where signals were transmitted by analogue and digital telephone channels. Using a baseline ASV system based on *joint factor analysis* (JFA), the work showed an increase in the EER of 1% to almost 70% when impostor accesses were replaced by replayed spoof attacks.

A physical access scenario was considered in [21]. While the baseline performance of the Gaussian mixture model- universal background model (GMM-UBM) ASV system was not reported, experiments showed that replay attacks produced a FAR of 93%.

The work in [18] introduced audio-visual spoofing (AVspoof) database for replay attack detection where the replayed signals are collected and played back using different low-quality (phones and laptop) and high-quality (laptop with loud speakers) devices. The study reported that FARs for replayed speech was 77.4% and 69.4%

for male and female, respectively, using a total variability system speaker recognition system. In this study, the EER for bona fide trials was 6.9% and 17.5% for those conditions. This study also includes presentation attack where speech signals created with voice conversion and speech synthesis were used in playback attack. In that case, higher FAR was observed, particularly when high-quality device is used for playback.

### 2.2.2 Countermeasures

A countermeasure for replay attack detection in the case of text-dependent ASV was reported in [5]. The approach is based upon the comparison of new access samples with stored instances of past accesses. New accesses which are deemed too similar to previous access attempts are identified as replay attacks. A large number of different experiments, all relating to a telephony scenario, showed that the countermeasures succeeded in lowering the EER in most of the experiments performed. While some form of text-dependent or challenge-response countermeasure is usually used to prevent replay attacks, text-independent solutions have also been investigated. The same authors in [20] showed that it is possible to detect replay attacks by measuring the channel differences caused by far-field recording [22]. While they show spoof detection error rates of less than 10% it is feasible that today's state-of-the-art approaches to channel compensation will render some ASV systems still vulnerable.

Two different replay attack countermeasures are compared in [21]. Both are based on the detection of differences in channel characteristics expected between licit and spoofed access attempts. Replay attacks incur channel noise from both the recording device and the loudspeaker used for replay and thus the detection of channel effects beyond those introduced by the recording device of the ASV system thus serves as an indicator of replay. The performance of a baseline GMM-UBM system with an EER of 40% under spoofing attack falls to 29% with the first countermeasure and a more respectable EER of 10% with the second countermeasure.

In another study [23], a speech database of 175 subjects has been collected for different kinds of replay attack. Other than the use of genuine voice samples for the legitimate speakers in playback, the voice samples recorded over the telephone channel were also used for unauthorised access. Further, a far-field microphone is used to collect the voice samples as eavesdropped (covert) recording. The authors proposed an algorithm motivated from music recognition system used for comparing recordings on the basis of the similarity of the local configuration of maxima pairs extracted from spectrograms of verified and reference recordings. The experimental results show the EER of playback attack detection to be as low as 1.0% on the collected data.

## 2.3 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 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 [24]: text analysis followed by 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 emerging end-to-end frameworks that generate speech waveforms directly from text inputs without using any additional modules.

Many approaches have been investigated, but there have been major paradigm shifts every ten years. 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 [25]. In the 1980s, the speech waveform generation component used a small database of phoneme units called *diphones* (the second half of one phoneme 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 [26]. In the 1990s, larger speech databases were collected and used to select more appropriate speech units that matched both phonemes and other linguistic contexts such as lexical stress and pitch accent in order to generate high-quality natural sounding synthetic speech with the appropriate prosody. This approach is generally referred to as *unit selection*, and is nowadays used in many speech synthesis systems [27, 28, 29, 30, 31].

In the late 2000s, several machine learning based data-driven approaches emerged. ‘Statistical parametric speech synthesis’ was one of the more popular machine learning approaches [32, 33, 34, 35]. In this approach, several acoustic parameters are modelled using a time-series stochastic generative model, typically a HMM. HMMs represent not only the phoneme sequences but also various contexts of the linguistic specification. Acoustic parameters generated from HMMs and selected according to the linguistic specification are then used to drive a vocoder, a simplified speech production model in which speech is represented by vocal tract parameters and excitation parameters in order to generate a speech waveform. HMM-based speech synthesisers [36, 37] can also learn speech models from relatively small amounts of 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) [38, 39].

In the 2010s, deep learning has significantly improved the performance of speech synthesis and led to a significant breakthrough. First, various types of deep neural networks are used to improve the prediction accuracy of the acoustic parameters [40, 41]. Investigated architectures include recurrent neural network [42, 43, 44], residual/highway network [45, 46], autoregressive network [47, 48], and generative adversarial networks (GAN) [49, 50, 51]. Furthermore, in the late 2010s conventional waveform generation modules that typically used signal processing and text analysis modules that used natural language processing were substituted by neural networks. This allows for neural networks capable of directly outputting the desired speech waveform samples from the desired text inputs. Successful architectures for direct waveform modelling include dilated convolutional autoregressive neural network, known as "Wavenet" [52] and hierarichical recurrent neural network, called "SampleRNN" [53]. Finally, we have also seen successful architectures that totally remove the hand-crafted linguistic features obtained through text analysis by relying in sequence-to-sequence systems. This system is called Tacotron [54]. As expected, the combination of these advanced models results in a very high-quality end-to-end TTS synthesis system [55, 56] and recent results reveal that the generated synthetic speech sounds as natural as human speech [56].

For more details and technical comparisons, please see the results of Blizzard Challenge, which annually compares the performance of speech synthesis systems built on the common database over decades [57, 58].

### 2.3.1 Spoofing

There is a considerable volume of research in the literature which has demonstrated the vulnerability of ASV to synthetic voices generated with a variety of approaches to speech synthesis. Experiments using formant, diphone, and unit-selection based synthetic speech in addition to the simple cut-and-paste of speech waveforms have been reported [19, 59, 20].

ASV vulnerabilities to HMM-based synthetic speech were first demonstrated over a decade ago [60] using an HMM-based, text-prompted ASV system [61] and an HMM-based synthesiser where acoustic models were adapted to specific human speakers [62, 63]. The ASV system scored feature vectors against speaker and background models composed of concatenated phoneme models. When tested with human speech, the ASV system achieved a FAR of 0% and a false rejection rate (FRR) of 7%. When subjected to spoofing attacks with synthetic speech, the FAR increased to over 70%, however, this work involved only 20 speakers.

Larger scale experiments using the Wall Street Journal corpus containing in the order of 300 speakers and two different ASV systems (GMM-UBM and SVM using Gaussian supervectors) was reported in [64]. Using an HMM-based speech synthesiser, the FAR was shown to rise to 86% and 81% for the GMM-UBM and SVM systems respectively representing a genuine threat to ASV. Spoofing experiments using HMM-based synthetic speech against a forensics speaker verification tool *BATVOX* was also reported in [65] with similar findings. Therefore, the above speech synthesisers were chosen as one of spoofing methods in the ASVspoof 2015 database.

Spoofing experiments using the above advanced DNNs or using spoofing-specific strategies such as GAN have not yet been properly investigated. Only a relatively small-scale spoofing experiment against a speaker recognition system using Wavenet, SampleRNN and GAN is reported in [66].

### 2.3.2 Countermeasures

Only a small number of attempts to discriminate synthetic speech from natural speech had been investigated before the ASVspoof challenge started. Previous work has demonstrated the successful detection of synthetic speech based on prior knowledge of the acoustic differences of specific speech synthesizers, such as the dynamic ranges of spectral parameters at the utterance level [67] and variance of higher order parts of mel-cepstral coefficients [68].

There are some attempts which focus on acoustic differences between vocoders and natural speech. Since the human auditory system is known to be relatively insensitive to phase [69], vocoders are typically based on a minimum-phase vocal tract model. This simplification leads to differences in the phase spectra between human and synthetic speech, differences which can be utilised for discrimination [64, 70].

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 [71, 72]. 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.

After the ASVspoof challenges took place, various types of countermeasures that work for both speech synthesis and voice conversion have been proposed. Please read the next section for the details of the recently developed countermeasures.

## *2.4 Voice conversion*

, in short, VC , is a spoofing attack against automatic speaker verification using an attackers natural voice which is converted towards that of the target. It aims to convert one speaker’s voice towards that of another and is a sub-domain of voice transformation [73]. Unlike TTS, which requires text input, voice conversion operates directly on speech inputs. However, speech waveform generation modules such as vocoders, may be the same as or similar to those for TTS.

A major application of VC is to personalise and create new voices for TTS synthesis systems and spoken dialogue systems. Other applications include speaking aid devices that generate more intelligible voice sounds to help people with speech disorders, movie dubbing, language learning, and singing voice conversion. The field has also attracted increasing interest in the context of ASV vulnerabilities for almost two decades [74].

Most voice conversion approaches require a parallel corpus where source and target speakers read out identical utterances and adopt a training phase which typically requires frame- or phone-aligned audio pairs of the source and target utterances and estimates transformation functions that convert acoustic parameters of the source speaker to those of the target speaker. This is called "parallel voice conversion". Frame alignment is traditionally achieved using dynamic time warping (DTW) on the source-target training audio files. Phone alignment is traditionally achieved using *automatic speech recognition* (ASR) and phone-level forth alignment. The estimated conversion function is then applied to any new audio files uttered by the source speaker [75].

A large number of estimation methods for the transformation functions have been reported starting in the late 1980s. In the late 1980's and 90's, simple techniques employing vector quantisation (VQ) with codebooks [76] or segmental codebooks [77] of paired source-target frame vectors were proposed to represent the transformation functions. However, these VQ methods introduced frame-to-frame discontinuity problems.

In the late 1990s and 2000s, *joint density Gaussian mixture model* (JDGMM) based transformation methods [78, 79] were proposed and have since then been actively improved by many researchers [80, 81]. This method still remains popular even now. Although this method achieves smooth feature transformations using a locally linear transformation, this method also has several critical problems such as over-smoothing [82, 83, 84] and over-fitting [85, 86] which leads to muffled quality of speech and degraded speaker similarity.

Therefore, in the early 2010, several alternative linear transformation methods were developed. Examples are partial least square (PLS) regression [85], tensor representation [87], a trajectory HMM [88], mixture of factor analysers [89], local linear transformation [82] or noisy channel models [90].

In parallel to the linear-based approaches, there have been studies on non-linear transformation functions such as support vector regression [91], kernel partial least square [92], and conditional restricted Boltzmann machines [93], neural networks [94, 95], highway network [96], and RNN [97, 98]. Data-driven frequency warping techniques [99, 100, 101] have also been studied.

Recently, deep learning has changed the above standard procedures for voice conversion and we can see many different solutions now. For instance, variational auto-encoder or sequence-to-sequence neural networks enable us to build VC systems without using frame level alignment [102, 103]. It has also been showed that a cycle-consistent adversarial network called "CycleGAN" [104] is one possible solution for building VC systems without using a parallel corpus. Wavenet can also be used as a replacement for the purpose of generating speech waveforms from converted acoustic features [105].

The approaches to voice conversion considered above are usually applied to the transformation of spectral envelope features, though the conversion of prosodic features such as fundamental frequency [106, 107, 108, 109] and duration [107, 110] has also been studied.

For more details and technical comparisons, please see results of Voice Conversion Challenges that compare the performance of VC systems built on a common database [111, 112].

### 2.4.1 Spoofing

When applied to spoofing, the aim with voice conversion is to synthesise a new speech signal such that the extracted ASV features are close in some sense to the target speaker. Some of the first works relevant to text-independent ASV spoofing were reported in [113, 114]. The work in [113] showed that baseline EER increased from 16% to 26% thanks to a voice conversion system which also converted prosodic aspects not modeled in typical ASV systems. This work targeted the conversion of spectral-slope parameters and showed that the baseline EER of 10% increased to over 60% when all impostor test samples were replaced with converted voices. Moreover, signals subjected to voice conversion did not exhibit any perceivable artefacts indicative of manipulation.

The work in [115] investigated ASV vulnerabilities to voice conversion based on JDGMMs [78] which requires a parallel training corpus for both source and target speakers. Even if the converted speech could be easily detectable by human listeners, experiments involving five different ASV systems showed their universal susceptibility to spoofing. The FAR of the most robust, JFA system increased from 3% to over 17%. Instead of vocoder-based waveform generation, unit selection approaches can be applied directly to feature vectors coming from the target speaker to synthesise converted speech [116]. Since they use target speaker data directly, unit-selection approaches arguably pose a greater risk to ASV than statistical approaches [117]. In the ASVspoof 2015 challenge, we therefore had chosen these popular VC methods as spoofing methods.

Other work relevant to voice conversion includes attacks referred to as artificial signals. It was noted in [118] that certain short intervals of converted speech yield extremely high scores or likelihoods. Such intervals are not representative of intelligible speech but they are nonetheless effective in overcoming typical ASV systems which lack any form of speech quality assessment. The work in [118] showed that artificial signals optimised with a genetic algorithm provoke increases in the EER from 10% to almost 80% for a GMM-UBM system and from 5% to almost 65% for a factor analysis (FA) system.

### 2.4.2 Countermeasures

Here, we provide an overview of countermeasure methods developed for the VC attacks before the ASVspoof challenge began.

Some of the first works to detect converted voice draws on related work in synthetic speech detection [119]. In [70, 120], cosine phase and modified group delay function (MGDF) based countermeasures were proposed. These are effective in detecting converted speech using vocoders based on minimum phase. In VC, it is, however, possible to use natural phase information extracted from a source speaker [114]. In this case, they are unlikely to detect converted voice.

Two approaches to artificial signal detection are reported in [121]. Experimental work shows that supervector-based SVM classifiers are naturally robust to such attacks, and that all the spoofing attacks they used could be detected by using an utterance-level variability feature, which detected the absence of the natural and dynamic variabilities characteristic of genuine speech. A related approach to detect converted voice is proposed in [122]. Probabilistic mappings between source and target speaker models are shown to typically yield converted speech with less short-term variability than genuine speech. Therefore, the thresholded, average pair-wise distance between consecutive feature vectors was used to detect converted voice with an EER of under 3%.

Due to fact that majority of VC techniques operate at the short-term frame level, more sophisticated long-term features such as temporal magnitude and phase modulation feature can also detect converted speech [123]. Another experiment reported in [124] showed that local binary pattern analysis of sequences of acoustic vectors can also be used for successfully detecting frame-wise JDGMM-based converted voice. However, it is unclear whether these features are effective in detecting recent VC systems that consider long-term dependency such as recurrent or autoregressive neural network models.

After the ASVspoof challenges took place, new countermeasures that works for both speech synthesis and voice conversion were proposed and evaluated. See the next section for a detailed review of the recently developed countermeasures.

## 3 Summary of the spoofing challenges

A number of independent studies confirm the vulnerability of ASV technology to spoofed voice created using voice conversion, speech synthesis, and playback [6]. Early studies on speaker anti-spoofing were mostly conducted on in-house speech corpora created using a limited number of spoofing attacks. The development of countermeasures using only a small number of spoofing attacks may not offer the generalisation ability in the presence of different or unseen attacks. There was a lack of publicly available corpora and evaluation protocol to help with comparing the results obtained by different researchers.

The [1] initiative aims to overcome this bottleneck by making available standard speech corpora consisting of a large number of spoofing attacks, evaluation protocols, and metrics to support a common evaluation and the benchmarking of different systems. The speech corpora were initially distributed by organising an evaluation challenge. In order to make the challenge simple and to maximise participation, the ASVspoof challenges so far involved only the detection of spoofed speech; in

[1] http://www.asvspoof.org/

effect, to determine whether a speech sample is genuine or spoofed. A training set and development set consisting of several spoofing attacks were first shared with the challenge participants to help them develop and tune their anti-spoofing algorithm. Next, the evaluation set without any label indicating genuine or spoofed speech was distributed, and the organisers asked the participants to submit scores within a specific deadline. Participants were allowed to submit scores of multiple systems. One of these systems was designated as the primary submission. Spoofing detectors for all primary submissions were trained using only the training data in the challenge corpus. Finally, the organisers evaluated the scores for benchmarks and ranking. The evaluation keys were subsequently released to the challenge participants. The challenge results were discussed with the participants in a special session in INTERSPEECH conferences, which also involved sharing knowledge and receiving useful feedback. To promote further research and technological advancements, the datasets used in the challenge are made publicly available.

The ASVspoof challenges have been organised twice so far. The first was held in 2015 and the second in 2017. A summary of the speech corpora used in the two challenges are shown in Table 1. In both the challenges, EER metric was used to evaluate the performance of spoofing detector. The EER is computed by considering the scores of genuine files as positive scores and those of spoofed files as negative scores. A lower EER means more accurate spoofing countermeasures. In practice, the EER is estimated using a specific *receiver operating characteristics convex hull* (ROCCH) technique with an open-source implementation[2] originating from outside the ASVspoof consortium. In the following subsections, we briefly discuss the two challenges. For more interested readers, [125] contains details of the 2015 edition while [126] discusses the results of the 2017 edition.

## *3.1 ASVspoof 2015*

The first ASVspoof challenge involved detection of artificial speech created using a mixture of voice conversion and speech synthesis techniques [125]. The dataset was generated with ten different artificial speech generation algorithms. The was based upon a larger collection spoofing and anti-spoofing (SAS) corpus (v1.0) [127] that consists of both natural and artificial speech. Natural speech was recorded from 106 human speakers using a high-quality microphone and without significant channel or background noise effects. In a speaker disjoint manner, the full database was divided into three subsets called the training, development, and evaluation set. Five of the attacks (S1-S5), named as *known attacks*, were used in the training and development set. The other five attacks, S6-S10, called *unknown attacks*, were used only in the evaluation set, along with the known attacks. Thus, this provides the possibility of assessing the generalisability of the spoofing detectors. The detailed evaluation plan is available in [128], describing the speech corpora and challenge rules.

[2] `https://sites.google.com/site/bosaristoolkit/`

Table 1: Summary of the datasets used in ASVspoof challenges.

| | **ASVspoof 2015** [125] | **ASVspoof 2017** [126] |
|---|---|---|
| Theme | Detection of artificially generated speech | Detection of replay speech |
| Speech format | $F_s$ = 16 kHz, 16 bit PCM | $F_s$ = 16 kHz, 16 bit PCM |
| Natural speech | Recorded using high-quality microphone | Recorded using different smart phones |
| Spoofed speech | Created with seven VC and three SS methods | Collected 'in the wild' by crowdsourcing using different microphone and playback devices from diverse environments |
| Spoofing types in train/dev/eval | 5 / 5 / 10 | 3 / 10 / 57 |
| No of speakers in train/dev/eval | 25 / 35 / 46 | 10 / 8 / 24 |
| No of genuine speech files in train/dev/eval | 3750 / 3497 / 9404 | 1508 / 760 / 1298 |
| No of spoofed speech files in train/dev/eval | 12625 / 49875 / 184000 | 1508 / 950 / 12008 |

Ten different spoofing attacks used in the ASVspoof 2015 are listed below:-

- **S1**: a simplified frame selection (FS) based voice conversion algorithm, in which the converted speech is generated by selecting target speech frames.
- **S2**: the simplest voice conversion algorithm which adjusts only the first mel-cepstral coefficient (C1) in order to shift the slope of the source spectrum to the target.
- **S3**: a speech synthesis algorithm implemented with the HMM based speech synthesis system (HTS3) using speaker adaptation techniques and only 20 adaptation utterances.
- **S4**: the same algorithm as S3, but using 40 adaptation utterances.
- **S5**: a voice conversion algorithm implemented with the voice conversion toolkit and with the Festvox system[3].
- **S6**: a VC algorithm based on joint density Gaussian mixture models (GMMs) and maximum likelihood parameter generation considering global variance.
- **S7**: a VC algorithm similar to S6, but using line spectrum pair (LSP) rather than mel-cepstral coefficients for spectrum representation.
- **S8**: a tensor-based approach to VC, for which a Japanese dataset was used to construct the speaker space.
- **S9**: a VC algorithm which uses kernel-based partial least square (KPLS) to implement a non-linear transformation function.
- **S10**: an SS algorithm implemented with the open-source MARY text-to-tpeech system (MaryTTS)[4].

[3] `http://www.festvox.org/`

[4] `http://mary.dfki.de/`

Table 2: Performance of top five systems in ASVspoof 2015 challenge (ranked according to the average % EER for all attacks) with respective features and classifiers.

| System | Avg. EER for | | | System |
|---|---|---|---|---|
| Identifier | known | unknown | all | Description |
| A [129] | 0.408 | 2.013 | 1.211 | *Features:* mel-frequency cepstral coefficients (MFCC),<br>Cochlear filter cepstral coefficients plus instantaneous frequency (CFCCIF).<br>*Classifier:* GMM. |
| B [130] | 0.008 | 3.922 | 1.965 | *Features:* MFCC, MFPC,<br>cosine-phase principal coefficients (CosPhasePCs).<br>*Classifier:* Support vector machine (SVM) with i-vectors. |
| C [131] | 0.058 | 4.998 | 2.528 | *Feature:* DNN-based with filterbank output and their deltas as input.<br>*Classifier:* Mahalanobis distance on s-vectors. |
| D [132] | 0.003 | 5.231 | 2.617 | *Features:* log magnitude spectrum (LMS),<br>residual log magnitude spectrum (RLMS), group delay (GD),<br>modified group delay (MGD), instantaneous frequency derivative (IF),<br>baseband phase difference (BPD), and pitch synchronous phase (PSP).<br>*Classifier:* Multilayer perceptron (MLP). |
| E [133] | 0.041 | 5.347 | 2.694 | *Features:* MFCC, product spectrum MFCC (PS-MFCC),<br>MGD with and without energy, weighted linear prediction group delay<br>cepstral coefficients (WLP-GDCCs), and MFCC<br>cosine-normalised phase-based cepstral coefficients (MFCC-CNPCCs).<br>*Classifier:* GMM. |

More details of how the SAS corpus was generated can be found in [127].

The organisers also confirmed the vulnerability to spoofing by conducting speaker verification experiments with this data and demonstrating considerable performance degradation in the presence of spoofing. With a state-of-the-art probabilistic linear discriminant analysis (PLDA) based ASV system, it is shown that in presence of spoofing, the average EER for ASV increases from 2.30% to 36.00% for male and 2.08% to 39.53% for female [125]. This motivates the development of the anti-spoofing algorithm.

For ASVspoof 2015, the challenge evaluation metric was the average EER. It is computed by calculating EERs for each attack and then taking average. The dataset was requested by 28 teams from 16 countries, 16 teams returned primary submissions by the deadline. A total of 27 additional submissions were also received. Anonymous results were subsequently returned to each team, who were then invited to submit their work to the ASVspoof special session for INTERSPEECH 2015.

Table 2 shows the performance of the top five systems in the ASVspoof 2015 challenge. The best performing system [129] uses a combination of *mel cesptral* and *cochlear filter cepstral coefficients plus instantaneous frequency* features with GMM back-end. In most cases, the participants have used fusion of multiple feature based systems to get better recognition accuracy. Variants of cepstral features computed from the magnitude and phase of short-term speech are widely used for the detection of spoofing attacks. As a back-end, GMM was found to outperform more advanced classifiers like i-vectors, possibly due to the use of short segments of high-quality speech not requiring treatment for channel compensation and background noise reduction. All the systems submitted in the challenge are reviewed in more detail [134].

## *3.2 ASVspoof 2017*

The is the second automatic speaker verification antispoofing and countermeasures challenge. Unlike the 2015 edition that used very high-quality speech material, the 2017 edition aims to assess spoofing attack detection with ”out in the wild” conditions. It focuses exclusively on replay attacks. The corpus originates from the recent *text-dependent RedDots* corpus[5], whose purpose was to collect speech data over mobile devices, in the form of smartphones and tablet computers, by volunteers from across the globe.

The replayed version of the original *RedDots* corpus was collected through a crowdsourcing exercise using various replay configurations consisting of varied devices, loudspeakers, and recording devices, under a variety of different environments across four European countries within the EU Horizon 2020-funded OCTAVE project[6], (see [126]). Instead of covert recording, we made a “short-cut” and took the digital copy of the target speakers’ voice to create the playback versions. The collected corpus is divided into three subsets: for training, development, and evaluation. Details of each are presented in Table 1. All three subsets are disjoint in terms of speakers and data collection sites. The training and development subsets were collected at three different sites. The evaluation subset was collected at the same three sites and also included data from two new sites. Data from the same site include different recordings and replaying devices and from different acoustic environments. The evaluation subset contains data collected from 161 replay sessions in 62 unique replay configurations[7]. More details regarding replay configurations can be found in [126, 135].

The primary evaluation metric is “pooled” EER. In contrast to the ASVspoof 2015 challenge, the EER is computed from scores pooled across all the trial segments rather than condition averaging. A baseline[8] system based on common GMM back-end classifier with constant Q cepstral coefficient (CQCC) [136, 137] features was provided to the participants. This configuration is chosen as baseline as it has shown best recognition performance on ASVspoof 2015. The baseline is trained using either combined training and development data (B01) or training data (B02) alone. The baseline system does not involve any kind of optimisation or tuning with

[5] `https://sites.google.com/site/thereddotsproject/`

[6] `https://www.octave-project.eu/`

[7] A **replay configuration** refers to a unique combination of room, replay device and recording device while a **session** refers to a set of source files, which share the same replay configuration.

[8] See *Appendix A.2. Software packages*

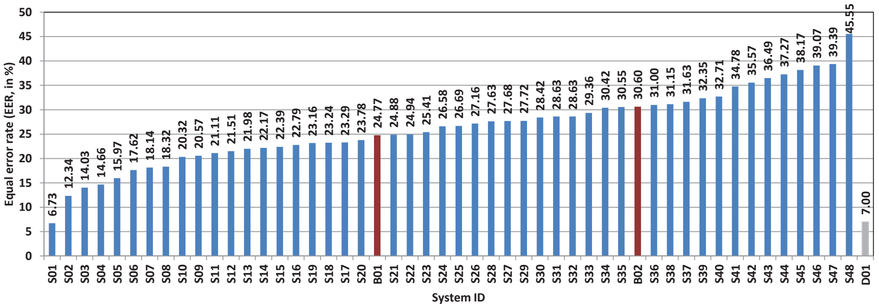

Fig. 3: Performance of the two baseline systems (B01 and B02) and the 49 primary systems (S01—S48 in addition to late submission D01) for the ASVspoof 2017 challenge. Results are in terms of the replay/non-replay EER (%).

respect to [136]. The dataset was requested by 113 teams, of which 49 returned primary submissions by the deadline. The results of the challenge were disseminated at a special session consisting of two slots at INTERSPEECH 2017.

Most of the systems are based on standard spectral features, such as CQCCs, MFCCs, and *perceptual linear prediction* (PLP). As a back-end, in addition to the classical GMM to model the replay and non-replay classes, it has also exploited the power of deep classifiers, such as *convolutional neural network* (CNN) or *recurrent neural network* (RNN). A fusion of multiple features and classifiers is also widely adopted by the participants. A summary of the top-10 primary systems is provided in Table 3. Results in terms of EER of the 49 primary systems and the baseline B01 and B02 are shown in Figure 3.

## 4 Advances in front-end features

The selection of appropriate features for a given classification problem is an important task. Even if the classic boundary to think between a feature extractor (front-end) and a classifier (back-end) as separate components is getting increasingly blurred with the use of end-to-end deep learning and other similar techniques, research on the ‘early’ components in a pipeline remains important. In the context of anti-spoofing for ASV, this allows the utilisation of one’s domain knowledge to guide the design of new discriminative features. For instance, earlier experience suggests that lack of spectral [70] and temporal [123] detail is characteristic of synthetic or voice-coded (vocoded) speech, and that low-quality replayed signals tend to experience loss of spectral details [143]. These initial findings sparked further research into developing advanced front-end features with improved robustness, generalisation across datasets, and other desideratum. As a matter of fact, in contrast to classic ASV (without spoofing attacks) where the most significant advancements have been in the back-end modelling [2], in ASV anti-spoofing, the features seem

Table 3: Summary of top 10 primary submissions to ASVspoof 2017. Systems' IDs are the same received by participants in the evaluation. The column 'Training' refers to the part of data used for training: train (T) and/or development (D).

| ID | Features | Post-proc. | Classifiers | Fusion | #Subs. | Training | Performances on eval subset (EER%) |
|---|---|---|---|---|---|---|---|
| S01 [138] | Log-power Spectrum, LPCC | MVN | CNN, GMM, TV, RNN | Score | 3 | T | 6.73 |
| S02 [139] | CQCC, MFCC, PLP | WMVN | GMM-UBM, TV-PLDA, GSV-SVM, GSV-GBDT, GSV-RF | Score | – | T | 12.34 |
| S03 | MFCC, IMFCC, RFCC, LFCC, PLP, CQCC, SCMC, SSFC | – | GMM, FF-ANN | Score | 18 | T+D | 14.03 |
| S04 | RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC | – | GMM | Score | 12 | T+D | 14.66 |
| S05 [140] | Linear filterbank feature | MN | GMM, CT-DNN | Score | 2 | T | 15.97 |
| S06 | CQCC, IMFCC, SCMC, Phrase one-hot encoding | MN | GMM | Score | 4 | T+D | 17.62 |
| S07 | HPCC, CQCC | MVN | GMM, CNN, SVM | Score | 2 | T+D | 18.14 |
| S08 [141] | IFCC, CFCCIF, Prosody | – | GMM | Score | 3 | T | 18.32 |
| S09 | SFFCC | No | GMM | None | 1 | T | 20.57 |
| S10 [142] | CQCC | – | ResNet | None | 1 | T | 20.32 |

to make the difference. In this section, we take a brief look at a few such methods emerging from the ASVspoof evaluations. The list is by no means exhaustive and the interested reader is referred to [134] for further discussion.

## *4.1 Front-ends for detection of voice conversion and speech synthesis spoofing*

The front-ends described below have been shown to provide good performance on the ASVspoof 2015 database of spoofing attacks based on voice conversion and speech synthesis. The first front-end was used in the ASVspoof 2015 challenge, while the rest were proposed later after the evaluation.

**Cochlear filter cepstral coefficients with instantaneous frequency (CFCCIF).** These features were introduced in [129] and successfully used as part of the top-ranked system in the ASVspoof 2015 evaluation. They combine cochlear filter cepstral coefficients (CFCC), proposed in [144], with instantaneous frequency [69]. CFCC are based on wavelet transform-like auditory transform and on some mechanisms of the cochlea of the human ear, such as hair cells and nerve spike density. To compute CFCC with instantaneous frequency (CFCCIF), the output of the nerve spike density envelope is multiplied by the instantaneous frequency, followed by the derivative operation and logarithm non-linearity. Finally, the discrete cosine transform (DCT) is applied to decorrelate the features and obtain a set of cepstral coefficients.

**Linear frequency cepstral coefficients (LFCC).** LFCCs are very similar to the widely used mel-frequency cepstral coefficients (MFCCs) [145], though the filters are placed in equal sizes for linear scale. This front-end is widely used in speaker recognition and has been shown to perform well in spoofing detection [146]. This technique performs a windowing on the signal, computes the magnitude spectrum using the short-time Fourier transform (STFT), followed by logarithm non-linearity and the application of a filterbank of linearly-spaced $N$ triangular filters to obtain a set of $N$ log-density values. Finally, the DCT is applied to obtain a set of cepstral coefficients.

**Constant Q cepstral coefficients (CQCC)**. This feature was proposed in [136, 137] for spoofing detection and it is based on the constant Q transform (CQT) [147]. The CQT is an alternative time-frequency analysis tool to the STFT that provides variable time and frequency resolution. It provides greater frequency resolution at lower frequencies but greater time resolution at higher frequencies. Figure 4 illustrates the extraction process. The CQT spectrum is obtained, followed by logarithm non-linearity and by a linearisation of the CQT geometric scale. Finally, cepstral coefficients are obtained though the DCT.

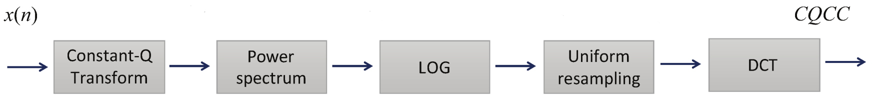

Fig. 4: Block diagram of CQCC feature extraction process.

As an alternative to CQCC, infinite impulse response constant-Q transform cepstrum (ICQC) features [148] use the infinite impulse response - constant Q transform [149], an efficient constant Q transform based on the IIR filtering of the fast Fourier transform (FFT) spectrum. It delivers multiresolution time-frequency analysis in a linear scale spectrum which is ready to be coupled with traditional cepstral analysis. The IIR-CQT spectrum is followed by the logarithm and decorrelation, either through the DCT or principal component analysis.

**Deep features for spoofing detection.** All of the above three features sets are hand-crafted and consists of a fixed sequence of standard digital signal processing operations. An alternative approach, seeing increased popularity across different machine learning problems, is to learn the feature extractor from a given data by using deep learning techniques [150, 151]. In speech-related applications, these features are widely employed for improving recognition accuracy [152, 153, 154]. The work in [155] uses deep neural network to generate bottleneck features for spoofing detection; that is, the activations of a hidden layer with a relatively small number of nodes compared to the size of other layers. The study in [156] investigates various features based on deep learning techniques. Different feed-forward DNNs are used to obtain frame-level deep features. Input acoustic features consisting of filterbank outputs with their first derivatives are used to train the network to discriminate between the natural and spoofed speech classes, and output of hidden layers are taken as deep features which are then averaged to obtain an utterance-level descriptor. RNNs are also proposed to estimate utterance-level features from input sequences of acoustic features. In another recent work [157], the authors have investigated deep features based on filterbank trained with the natural and artificial speech data. A feed forward neural network architecture called here as filterbank neural network (FBNN) is used here that includes a linear hidden layer, a sigmoid hidden layer and a softmax output layer. The number of nodes in the output is six; and of them, five are for the number of spoofed classes in the training set, and the remaining one is for natural speech. The filterbanks are learned using the stochastic gradient descent algorithm. The cepstral features extracted using these DNN-based features are shown to be better than the hand-crafted cepstral coefficients.

**Scattering cepstral coefficients.** This feature for spoofing detection was proposed in [158]. It relies upon *scattering spectral decomposition* [159, 160]. This transform is a hierarchical spectral decomposition of a signal based on wavelet filterbanks (constant Q filters), modulus operator, and averaging. Each level of decomposition processes the input signal (either the input signal for the first level of decomposition, or the output of a previous level of decomposition) through the wavelet filterbank and takes the absolute value of filter outputs, producing a scalogram. The scattering coefficients at a certain level are estimated by windowing the scalogram signals and computing the average value within these windows. A two-level scattering decomposition has been shown to be effective for spoofing detection [158]. The final feature vector is computed by taking the DCT of the vector obtained by concatenating the logarithms of the scattering coefficients from all levels and retaining the first a few coefficients. The “interesting” thing about scattering transform is its stability to small signal deformation and more details of the temporal envelopes than MFCCs [159, 158].

**Fundamental frequency variation features.** The prosodic features are not as successful as cepstral features in detecting artificial speech on ASVspoof 2015, though some earlier results on PAs indicate that pitch contours are useful for such tasks [6]. In a recent work [161], the author use fundamental frequency variation (FFV) for this. The FFV captures pitch variation at the frame-level and provides complementary information on cepstral features [162]. The combined system gives a very promising performance for both known and unknown conditions on ASVspoof evaluation data.

**Phase-based features.** The phase-based features are also successfully used in PAD systems for ASVspoof 2015. For example, relative phase shift (RPS) and modified group delay (MGD) based features are explored in [163]. The authors in [164] have investigated relative phase information (RPI) features. Though the performances on seen attacks are promising with these phase-based features, the performances noticeably degrade for unseen attacks, particularly for S10.

**General observations regarding front-ends for artificial speech detection.** Beyond the feature extraction method used, there are two general findings common to any front end [146, 129, 137, 148]. The first refers to the use of dynamic coefficients. The first and second derivatives of the static coefficients, also known as velocity and acceleration coefficients, respectively are found important to achieve good spoofing detection performance. In some cases, the use of only dynamic features is superior to the use of static plus dynamic coefficients [146]. This is not entirely surprising, since voice conversion and speech synthesis techniques may fail to model the dynamic properties of the speech signals, introducing artefacts that help the discrimination of spoofed signals. The second finding refers to the use of speech activity detection. In experiments with ASVspoof 2015 corpus, it appears that the silence regions also contain useful information for discriminating between natural and synthetic speech. Thus, retaining non-speech frames turns out to be a better choice for this corpus [146]. This is likely due to the fact that non-speech regions are usually replaced with noise during the voice conversion or speech synthesis operation. However, this could be a database-dependent observation, thus detailed investigations are required.

## *4.2 Front-ends for replay attack detection*

The following front-ends have been proposed for the task of replay spoofing detection, and evaluated in replayed speech databases such as the BTAS 2016 and ASVspoof 2017. Many standard front-ends, such as MFCC, LFCC, and PLP, have been combined to improve the performance of replay attack detection. Other front-ends proposed for synthetic and converted speech detection (CFCCIF, CQCC) have been successfully used for the replay detection task. In general, and in opposition to the trend for synthetic and converted speech detection, the use of static coefficients has been shown to be crucial for achieving good performance. This may be explained by the nature of the replayed speech detection task, where detecting changes in the channel captured by static coefficients helps with the discrimination of natural and replayed speech. Two additional front-ends are described next.

**Inverted mel frequency cepstral coefficients (IMFCC).** This front-end is relatively simple and similar to the standard MFCC. The only difference is that the filterbank follows an inverted mel scale; that is, it provides an increasing frequency resolution (narrower filters) when frequency increases, and a decreased frequency resolution (wider filters) for decreasing frequency, unlike the mel scale [165]. This front-end was used as part of the top-ranked system of the Biometrics: Theory, Applications, and Systems (BTAS) 2016 speaker antispoofing competition [7].

**Features based on convolutional neural networks.** In the recent ASVspoof 2017 challenge, the use of deep learning frameworks for feature learning was proven to be key in achieving good replay detection performance. In particular, convolutional neural networks have been successfully used to learn high-level utterance-level features which can later be classified with simple classifiers. As part of the top-ranked system [138] in the ASVspoof 2017 challenge, a light convolutional neural network architecture [166] is fed with truncated normalised FFT spectrograms (to force fixed data dimensions). The network consists of a set of convolutional layers, followed by a fully-connected layer. The last layer contains two outputs with softmax activation corresponding to the two classes. All layers use the max-feature-map activation function [166], which acts as a feature selector and reduces the number of feature maps by half on each layer. The network is then trained to discriminate between the natural and spoofed speech classes. Once the network is trained, it is used to extract a high-level feature vector which is the output of the fully connected layer. All the test utterances are processed to obtain high-level representations, which are later classified with an external classifier.

**Other hand-crafted features.** Many other features have also been used for replayed speech detection in the context of the ASVspoof 2017 database. Even if the performances of single systems using such features are not always high, they are shown to be complementary when fused at the score level [167], similar to conventional ASV research outside of the spoofing detection. These features include MFCC, IMFCC, rectangular filter cepstral coefficients (RFCCs), PLP, CQCC, spectral centroid magnitude coefficients (SCMC), subband spectral flux coefficient (SSFC), and variable length Teager energy operator energy separation algorithm-instantaneous frequency cosine coefficients (VESA-IFCC). Though, of course, one usually then has to further train the fusion system, which makes the system more involved concerning practical applications.

## 5 Advances in back-end classifiers

In the natural vs. spoof classification problem, two main families of approaches have been adopted, namely generative and discriminative. Generative approaches include those of GMM-based classifiers and i-vector representations combined with support vector machines (SVMs). As for discriminative approaches, deep learning based techniques have become more popular. Finally, new deep learning end-to-end solutions are emerging. Such techniques perform the typical pipeline entirely through deep learning, from feature representation learning and extraction to the final classification. While including such approaches into the traditional classifiers category may not be the most precise, they are included in this classifiers section for simplicity.

### *5.1 Generative approaches*

**Gaussian mixture model (GMM) classifiers.** Considering two classes, namely natural and spoofed speech, one GMM can be learned for each class using appropriate training data. In the classification stage, an input utterance is processed to obtain its likelihoods with respect to the natural and spoofed models. The resulting classification score is the log-likelihood ratio between the two competing hypotheses; in effect, those of the input utterance belonging to the natural and to the spoofed classes. A high score supports the former hypothesis, while a low score supports the latter. Finally, given a test utterance, classification can be performed by thresholding the obtained score. If the score is above the threshold, the test utterance is classified as natural, and otherwise, it is classified as spoof. Many proposed anti-spoofing systems use GMM classifiers [146, 129, 136, 168, 155, 158, 148].

**I-vector.** The state-of-the-art paradigm for speaker verification [169] has been explored for spoofing detection [170, 171]. Typically, an i-vector is extracted from an entire speech utterance and used as a low-dimensional, high-level feature which is later classified by means of a binary classifier, commonly cosine distance measure or support vector machine (SVM). Different amplitude- and phase-based frontends [130, 138] can be employed for the estimation of i-vectors. A recent work shows that data selection for i-vector extractor training (also known as **T** matrix) is an important factor for achieving completive recognition accuracy [172].

## 5.2 Discriminative approaches

**DNN classifiers.** Deep learning based classifiers have been explored for use in the task of natural and spoofed speech discrimination. In [173, 155], several front-ends are evaluated with neural network classifier consisting of several hidden layers with sigmoid nodes and softmax output, which is used to calculate utterance posteriors. However, the implementation detail of the DNNs - such the number of nodes, the cost function, the optimization algorithm and the activation functions - is not precisely mentioned in those work and the lack of this very relevant information make it difficult to reproduce the results.

In a recent work [174], a five-layer DNN spoofing detection system is investigated for ASVspoof 2015 which uses a novel scoring method, termed in the paper as *human log-likelihoods* (HLLs). Each of the hidden layers has 2048 nodes with a sigmoid activation function. The network has six softmax output layers. The DNN is implemented using a computational network toolkit[9] and trained with stochastic gradient descent methods with dynamics information of acoustic features, such as spectrum-based cepstral coefficients (SBCC) and CQCC as input. The cross entropy function is selected as the cost function and the maximum training epoch is chosen as 120. The mini-batch size is set to 128. The proposed method shows considerable PAD detection performance. The author obtain an EER for S10 of 0.255% and average EER for all attacks of 0.045% when used with CQCC acoustic features. These are the best reported performance in ASVspoof 2015 so far.

**DNN-based end-to-end approaches.** End-to-end systems aim to perform all the stages of a typical spoofing detection pipeline, from feature extraction to classifi-

[9] `https://github.com/Microsoft/CNTK`

cation, by learning the network parameters involved in the process as a whole. The advantage of such approaches is that they do not explicitly require prior knowledge of the spoofing attacks as required for the development of acoustic features. Instead, the parameters are learned and optimised from the training data. In [175], a convolutional long short-term memory (LSTM) deep neural network (CLDNN) [176] is used as an end-to-end solution for spoofing detection. This model receives input in the form of a sequence of raw speech frames and outputs a likelihood for the whole sequence. The CLDNN performs time-frequency convolution through CNN to reduce spectral variance, long-term temporal modelling by using a LSTM, and classification using a DNN. Therefore, it is a entirely an end-to-end solution which does not rely on any external feature representation. The works in [177, 138] propose other end-to-end solutions by combining convolutional and recurrent layers, where the first act as a feature extractor and the second models the long-term dependencies and acts as a classifier. Unlike the work in [175], the input data is the FFT spectrogram of the speech utterance and not the raw speech signal. In [178], the authors have investigated CNN-based end-to-end system for PAD where the raw speech is used to jointly learn the feature extractor and classifier. Score-level combination of this CNN system with standard long-term spectral statistics based system shows considerable overall improvement.

## 6 Other PAD approaches

While most of the studies in voice PAD detection research focus on algorithmic improvements for discriminating natural and artificial speech signals, some recent studies have explored utilising additional information collected using special additional hardware to protect ASV system from presentation attacks [179, 180, 181, 182]. Since an intruder can easily collect voice samples for the target speakers using covert recording; the idea there is to detect and recognise supplementary information related to the speech production process. Moreover, by its nature, that supplementary information is difficult, if not impossible, to mimic using spoofing methods in the practical scenario. These PAD techniques have shown excellent recognition accuracy in the spoofed condition, at the cost of additional setup in the data acquisition step.

The work presented in [180, 181] utilises the phenomenon of , which is a distortion in human breath when it reaches a microphone [183]. During natural speech production, the interactions between the airflow and the vocal cavities may result in a sort of plosive burst, commonly know as pop noise, which can be captured via a microphone. In the context of professional audio and music production, pop noise is unwanted and is eliminated during the recording or mastering process. In the context of ASV, however, it can help in the process of PAD. The basic principle is that a replay sound from a loudspeaker does not involve the turbulent airflow generating the pop noise as in the natural speech. The authors in [180, 181] have developed a pop noise detector which eventually distinguishes natural speech from playback recording as well as synthetic speech generated using VC and SS methods. In experiments with 17 female speakers, a tandem detection system that combines both single- and double-channel pop noise detection gives the lowest ASV error rates in the PA condition.

The authors in [179] have introduced the use of a smartphone-based *magnetometer* to detect voice presentation attack. The conventional loudspeakers, which are used for playback during access of the ASV systems, generate sound using acoustic transducer and generate a magnetic field. The idea, therefore, is to capture the use of loudspeaker by sensing the magnetic field which would be absent from human vocals. Experiments were conducted using playback from 25 different conventional loudspeakers, ranging from low-end to high-end and placed in different distances from the smartphone that contains the ASV system. A speech corpus of five speakers was collected for the ASV experiments executed using an open-source ASV toolkit, SPEAR[10]. Experiments were conducted with other datasets, using a similarly limited number of speakers. The authors demonstrated that the magnetic field based detection can be reliable for the detection of playback within 6-8 cm from the smartphone. They further developed a mechanism to detect the size of the sound source to prevent the use of small speakers, such as ear phones.

The authors in [184, 185] utilise certain acoustics concepts to prevent ASV systems from PAs. They first introduced a method [184] that estimates dynamic sound source position (articulation position within mouth) of some speech sounds using a small array using *microelectromechanical systems* (MEMS) microphones embedded in mobile devices and compare it with loudspeakers, which have a flat sound source. In particular, the idea is to capture the dynamics of *time-difference-of-arrival* (TDOA) in a sequence of speech sounds to the microphones of the smartphone. Such unique TDOA changes, which do not exist under replay conditions, are used for detecting replay attacks. The similarities between the TDOAs of test speech and user templates are measured using probability function under Gaussian assumption and correlation measure as well as their combinations. Experiments involving 12 speakers and three different types of smartphone demonstrate a low EER and high PAD accuracy. The proposed method is seen to remain robust despite the change of smartphones during the test and the displacements.

In [185], the same research group has used the idea of the *Doppler effect* to detect the replay attack. The idea here is to capture the *articulatory gestures* of the speakers when they speak a pass-phrase. The smartphone acts as a Doppler radar and transmits a high frequency tone at 20 kHz from the built-in speaker and senses the reflections using the microphone during authentication process. The movement of the speaker's articulators during vocalisation creates a speaker-dependent Doppler frequency shift at around 20 kHz, which is stored along with the speech signal during the speaker-enrolment process. During a playback attack, the Doppler frequency shift will be different due to the lack of articulatory movements. Energy-based frequency features and frequency-based energy features are computed from a band of 19.8 kHz and 20.2 kHz. These features are used to discriminate between the natu-

[10] `https://www.idiap.ch/software/bob/docs/bob/bob.bio.spear/stable/index.html`

ral and replayed voice; and the similarity scores are measured in terms of Pearson correlation coefficient. Experiments are conducted with a dataset of 21 speakers and using three different smartphones. The data also includes test speech for replay attack with different loudspeakers and for impersonation attack with four different impersonators. The proposed system was demonstrated to be effective in achieving low EER for both types of attacks. Similar to [184], the proposed method indicated robustness to the phone placement.

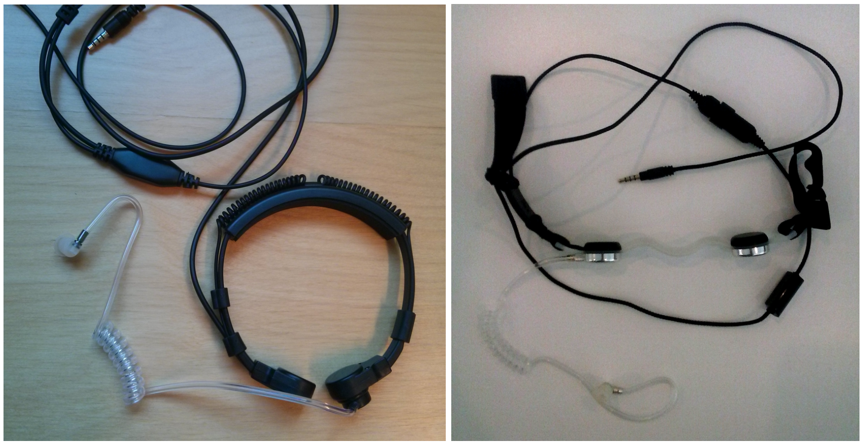

Fig. 5: Throat-microphones used in [182] [Reprinted with permission from IEEEACM Transactions on (T-ASL) Audio, Speech, and Language Processing].

The work in [182] introduces the use of a specific non-acoustic sensor, *throat microphone* (TM), or laryngophone, to enhance the performance of the voice PAD system. An example of such microphones is shown in Fig. 5. The TM is used with a conventional acoustic microphone (AM) in a dual-channel framework for robust speaker recognition and PAD. Since this type of microphone is attached to the speaker's neck, it would be difficult for the attacker to obtain a covert recording of the target speaker's voice. Therefore, one possibility for the intruder is to use the stolen recording from an AM and to try to record it back using a TM for accessing the ASV system. A speech corpus of 38 speakers was collected for the ASV experiments. The dual-channel setup yielded considerable ASV for both licit and spoofed conditions. The performance is further improved when this ASV system is integrated with the dual-channel based PAD. The authors show zero FAR for replay imposters by decision fusion of ASV and PAD.

All of the above new PAD methods deviating from the "mainstream" of PAD research in ASV are reported to be reliable and useful in specific application scenarios for identifying presentation attacks. The methods are also fundamentally different and difficult to compare in the same settings. Since the authors focus on the methodological aspects, experiments are mostly conducted on a dataset of limited number of speakers. Extensive experiments with more subjects from diverse environmental conditions should be performed to assess their suitability for real-world deployment.

## 7 Future directions of anti-spoofing research

The research in ASV anti-spoofing is becoming popular and well-recognised in the speech processing and voice-biometric community. The state-of-the-art spoofing detector gives promising accuracy in the benchmarking of spoofing countermeasures. Further work is needed to address a number of specific issues regarding its practical use. A number of potential topics for consideration in further work are now discussed.

- **Noise, reverberation and channel effect.** Recent studies indicate that spoofing countermeasures offer little resistance to additive noise [186, 187], reverberation [188] and channel effect [189] even though their performances on "clean" speech corpus are highly promising. The relative degradation of performance is actually much worse than the degradation of a typical ASV system under the similar mismatch condition. One reason could be that, at least until the ASVspoof 2017 evaluation, the methodology developed has been driven in clean, high-quality speech. In other words, the community might have developed its methods implicitly for laboratory testing. The commonly used speech enhancement algorithms also fail to reduce the mismatch due to environmental differences, though multi-condition training [187] and more advanced training methods [190] have been found useful. The study presented in [189] shows considerable degradation of PAD performance even in *matched* acoustic conditions. The feature settings used for the original corpus gives lower accuracy when both training and test data are digitally processed with the telephone channel effect. These are probably because the spoofing artefacts themselves act as extrinsic variabilities which degrade the speech quality in some way. Since the task of spoofing detection is related to detecting those artefacts, the problem becomes more difficult in the presence of small external effects due to variation in environment and channel. These suggests further investigations need to be carried out for the development of robust spoofing countermeasures.
- **Generalisation of spoofing countermeasures.** The property of spoofing countermeasures for detecting new kinds of speech presentation attack is an important requirement for their application in the wild. Study explores that countermeasure methods trained with a class of spoofing attacks fail to generalise this for other classes of spoofing attack [191, 167]. For example, PAD systems trained with VC and SS based spoofed speech give a very poor performance for playback detection [192]. The results of the first two ASVspoof challenges also reveal that detecting the converted speech created with an "unknown" method or the playback voice recording in a new replay session are difficult to detect. These clearly indicate the overfitting of PAD systems with available training data. Therefore, further investigation should be conducted to develop attack-

independent universal spoofing detector. Other than the unknown attack issue, generalisation is also an important concern for cross-corpora evaluation of the PAD system [193]. This specific topic is discussed in chapter 19 of this book.

- **Investigations with new spoofing methods.** The studies of converted spoof speech mostly focused on methods based on classical signal processing and machine learning techniques. Recent advancements in VC and SS research with deep learning technology show significant improvements in creating high quality synthetic speech [52]. The GAN [194] can be used to create (generator) spoofed voices with relevant feedback from the spoofing countermeasures (discriminator). Some preliminary studies demonstrate that the GAN-based approach can make speaker verification systems more vulnerable to presentation attacks [195, 66]. More detailed investigations should be conducted on this direction for the development of countermeasure technology to guard against this type of advanced attack.
- **Joint operations of PAD and ASV.** The ultimate goal of developing PAD system is to protect the recogniser, the ASV system from imposters with spoofed speech. So far, the majority of the studies focused on the evaluation of standalone countermeasures. The integration of these two systems is not trivial number of reasons. First, standard linear output score fusion techniques, being extensively used to combine homogenous ASV system, are not appropriate since the ASV and its countermeasures are trained to solve two different tasks. Second, an imperfect PAD can increase the false alarm rate by rejecting genuine access trials [196]. Thirdly, and more fundamentally, it is not obvious whether improvements in standalone spoofing countermeasures should improve the overall system as a whole: a nearly perfect PAD system with close to zero EER may fail to protect ASV system in practice if not properly calibrated [197]. In a recent work [198], the authors propose a modification in a GMM-UBM based ASV system to make it suitable for both licit and spoofed conditions. The joint evaluation of PAD and ASV, as well as their combination techniques, certainly deserves further attention. Among other feedback received from the attendees of the ASVspoof 2017 special session organised during INTERSPEECH 2017, it was proposed that the authors of this chapter consider shifting the focus from standalone spoofing to more ASV-centric solutions in future. We tend to agree. In our recent work [199], we propose a new cost function for joint assessment of PAD and ASV system. In another work [200], we propose a new fusion method for combining scores of countermeasures and recognisers. This work also explores speech features which can be used both for PAD and ASV.

## 8 Conclusion

This contribution provides an introduction to the different voice presentation attacks and their detection methods. It then reviews previous works with a focus on recent progress in assessing the performance of PAD systems. We have also briefly reviewed two recent ASVspoof challenges organised for the detection of voice PAs. This study includes discussion of recently developed features and the classifiers which are predominantly used in ASVspoof evaluations. We further include an extensive survey on alternative PAD methods. Apart from the conventional voice-based systems that use statistical properties of natural and spoofed speech for their discrimination, these recently developed methods utilise a separate hardware for the acquisition of other signals such as pop noise, throat signal, and extrasensory signals with smartphones for PAD. The current status of these non-mainstream approaches to PAD detection is somewhat similar to the status of the now more-or-less standard methods for artificial speech and replay PAD detection some three to four years ago: they are innovative and show promising results, but the pilot experiments have been carried out on relatively small and/or proprietary datasets, leaving an open question as to how scalable or generalisable these solutions are in practice. Nonetheless, in the long run and noting especially the rapid development of speech synthesis technology, it is likely that the quality of artificial/synthetic speech will eventually be indistinguishable from that of natural human speech. Such future spoofing attacks therefore could not be detected using the current mainstream techniques that focus on spectral or temporal details of the speech signal, but will require novel ideas that benefit from auxiliary information, rather than just the acoustic waveform.

In the past three years, the progress in voice PAD research has been accelerated by the development and free availability of speech corpus such as the ASVspoof series, SAS, BTAS 2016, AVSpoof. The work discussed several open challenges which show that this problem requires further attention to improving robustness due to mismatch condition, generalisation to new type of presentation attacks, and so on. Results from joint evaluations with integrated ASV system are also an important requirement for practical applications of PAD research. We think, however, that this extensive review will be of interest not only to those involved in voice PAD research but also to voice-biometrics researchers in general.

## Appendix A. Action towards reproducible research

### *A.1. Speech corpora*

1. Spoofing and Anti-Spoofing (SAS) database v1.0: This database presents the first version of a speaker verification spoofing and anti-spoofing database, named SAS corpus [201]. The corpus includes nine spoofing techniques, two of which are speech synthesis, and seven are voice conversion.
   Download link: `http://dx.doi.org/10.7488/ds/252`
2. ASVspoof 2015 database: This database has been used in the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015). Genuine speech is collected from 106 speakers (45 male, 61 female) and with no significant channel or background noise effects. Spoofed speech is gener-

ated from the genuine data using a number of different spoofing algorithms. The full dataset is partitioned into three subsets, the first for training, the second for development and the third for evaluation.
Download link: `http://dx.doi.org/10.7488/ds/298`

3. ASVspoof 2017 database: This database has been used in the Second Automatic Speaker Verification Spoofing and Countermeasuers Challenge: ASVspoof 2017. This database makes an extensive use of the recent text-dependent RedDots corpus, as well as a replayed version of the same data. It contains a large amount of speech data from 42 speakers collected from 179 replay sessions in 62 unique replay configurations.
Download link: `http://dx.doi.org/10.7488/ds/2313`

### *A.2. Software packages*

1. Feature extraction techniques for anti-spoofing: This package contains the MATLAB implementation of different acoustic feature extraction schemes as evaluated in [146].
Download link: `http://cs.joensuu.fi/~sahid/codes/AntiSpoofing_Features.zip`
2. Baseline spoofing detection package for ASVspoof 2017 corpus: This package contain the MATLAB implementations of two spoofing detectors employed as baseline in the official ASVspoof 2017 evaluation. They are based on constant Q cepstral coefficients (CQCC) [137] and Gaussian mixture model classifiers.
Download link: `http://audio.eurecom.fr/software/ASVspoof2017_baseline_countermeasures.zip`

## References

1. T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12 – 40, 2010.
2. J.H.L. Hansen and T. Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99, 2015.
3. ISO/IEC 30107. Information technology – biometric presentation attack detection. *International Organization for Standardization*, 2016.
4. T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A.K. Sarkar, N. Thomsen, V. Hautamäki, N. Evans, and Z.-H. Tan. Utterance verification for text-dependent speaker recognition: A comparative assessment using the reddots corpus. In *Proc. Interspeech*, pages 430–434, 2016.
5. W. Shang and M. Stevenson. Score normalization in playback attack detection. In *Proc. ICASSP*, pages 1678–1681. IEEE, 2010.
6. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66(0):130 – 153, 2015.
7. P. Korshunov, S. Marcel, H. Muckenhirn, A.R. Gonçalves, A.G.S. Mello, R.P.V. Violato, F.O. Simoes, M.U. Neto, M. de A. Angeloni, J.A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul,

G. Saha, and M. Sahidullah. Overview of BTAS 2016 speaker anti-spoofing competition. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2016.

8. N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon. Speaker recognition anti-spoofing. In S. Marcel, S. Z. Li, and M. Nixon, editors, *Handbook of biometric anti-spoofing*. Springer, 2014.
9. S. Marcel, S. Z. Li, and M. Nixon, editors. *Handbook of biometric anti-spoofing: trusted biometrics under spoofing attacks*. Springer, 2014.
10. M. Farrús Cabeceran, M. Wagner, D. Erro, and H. Pericás. Automatic speaker recognition as a measurement of voice imitation and conversion. *The Intenational Journal of Speech. Language and the Law*, 1(17):119–142, 2010.
11. P. Perrot, G. Aversano, and G. Chollet. Voice disguise and automatic detection: review and perspectives. *Progress in nonlinear speech processing*, pages 101–117, 2007.
12. E. Zetterholm. Detection of speaker characteristics using voice imitation. In *Speaker Classification II*, pages 192–205. Springer, 2007.
13. Y.W. Lau, M. Wagner, and D. Tran. Vulnerability of speaker verification to voice mimicking. In *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pages 145–148. IEEE, 2004.
14. Y.W. Lau, D. Tran, and M. Wagner. Testing voice mimicry with the YOHO speaker verification corpus. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 15–21. Springer, 2005.
15. J. Mariéthoz and S. Bengio. Can a professional imitator fool a GMM-based speaker verification system? Technical report, IDIAP, 2005.
16. S. Panjwani and A. Prakash. Crowdsourcing attacks on biometric systems. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 257–269, 2014.
17. R.G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, 72:13–31, 2015.
18. S.K. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel. On the vulnerability of speaker verification to realistic voice spoofing. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, 2015.
19. J. Lindberg and M. Blomberg. Vulnerability in speaker verification-a study of technical impostor techniques. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1211–1214, 1999.
20. J. Villalba and E. Lleida. Speaker verification performance degradation against spoofing and tampering attacks. In *FALA 10 workshop*, pages 131–134, 2010.
21. Z. F. Wang, G. Wei, and Q. H. He. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *2011 International Conference on Machine Learning and Cybernetics*, volume 4, pages 1708–1713, 2011.
22. J. Villalba and E. Lleida. Preventing replay attacks on speaker verification systems. In *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, pages 1–8. IEEE, 2011.
23. J. Gałka, M. Grzywacz, and R. Samborski. Playback attack detection for text-dependent speaker verification over telephone channels. *Speech Communication*, 67:143–153, 2015.
24. P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.
25. D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995, 1980.
26. E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.*, 9:453–467, 1990.
27. A. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, pages 373–376, 1996.
28. A. Breen and P. Jackson. A phonologically motivated method of selecting nonuniform units. In *Proc. ICSLP*, pages 2735–2738, 1998.
29. R. E. Donovan and E. M. Eide. The IBM trainable speech synthesis system. In *Proc. ICSLP*, pages 1703–1706, 1998.

30. B. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS system. In *Proc. Joint ASA, EAA and DAEA Meeting*, pages 15–19, 1999.
31. G. Coorman, J. Fackrell, P. Rutten, and B. Coile. Segment selection in the L & H realspeak laboratory TTS system. In *Proc. ICSLP*, pages 395–398, 2000.
32. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2347–2350, 1999.
33. Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang. USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Proc. the Blizzard Challenge Workshop*, 2006.
34. A.W. Black. CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Proc. Interspeech*, pages 1762–1765, 2006.
35. H. Zen, T. Toda, M. Nakamura, and K. Tokuda. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. Syst.*, E90-D(1):325–333, 2007.
36. H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, November 2009.
37. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Speech, Audio & Language Process.*, 17(1):66–83, Jan. 2009.
38. C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.*, 9:171–185, 1995.
39. P. C. Woodland. Speaker adaptation for continuous density HMMs: A review. In *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, page 119, 2001.
40. H. Ze, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, pages 7962–7966, May 2013.
41. Z. H. Ling, L. Deng, and D. Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2129–2139, Oct 2013.
42. Y. Fan, Y. Qian, F.-L. Xie, and F.K. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proc. Interspeech*, pages 1964–1968, September 2014.
43. H. Zen and H. Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. ICASSP*, pages 4470–4474, April 2015.
44. Z. Wu and S. King. Investigating gated recurrent networks for speech synthesis. In *Proc. ICASSP*, pages 5140–5144, March 2016.
45. X. Wang, S. Takaki, and J. Yamagishi. Investigating very deep highway networks for parametric speech synthesis. In *9th ISCA Speech Synthesis Workshop*, pages 166–171, September 2016.
46. X. Wang, S. Takaki, and J. Yamagishi. Investigating very deep highway networks for parametric speech synthesis. *Speech Communication*, 96:1 – 9, 2018.
47. X. Wang, S. Takaki, and J. Yamagishi. An autoregressive recurrent mixture density network for parametric speech synthesis. In *Proc. ICASSP*, pages 4895–4899, March 2017.
48. X. Wang, S. Takaki, and J. Yamagishi. An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis. In *Proc. Interspeech*, pages 1059–1063, September 2017.
49. Y. Saito, S. Takamichi, and H. Saruwatari. Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis. In *Proc. ICASSP*, pages 4900–4904, 2017.
50. Y. Saito, S. Takamichi, and H. Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):84–96, Jan 2018.
51. T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *Proc. ICASSP*, pages 4910–4914, 2017.

52. A. Van D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
53. S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.
54. Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010, 2017.
55. A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in Neural Information Processing Systems*, pages 2966–2974, 2017.
56. J. Shen, M. Schuster, N. Jaitly, R.J. Skerry-Ryan, R.A. Saurous, R.J. Weiss, R. Pang, Y. Agiomyrgiannakis, Y. Wu, Y. Zhang, Y. Wang, Z. Chen, and Z. Yang. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. ICASSP*, 2018.
57. S. King. Measuring a decade of progress in text-to-speech. *Loquens*, 1(1):006, 2014.
58. S. King, L. Wihlborg, and W. Guo. The blizzard challenge 2017. In *Proc. Blizzard Challenge Workshop, Stockholm, Sweden*, 2017.
59. F.H. Foomany, A. Hirschfield, and M. Ingleby. Toward a dynamic framework for security evaluation of voice verification systems. In *Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, pages 22–27, 2009.
60. T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi. On the security of HMM-based speaker verification systems against imposture using synthetic speech. In *Proc. EUROSPEECH*, 1999.
61. T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Commun.*, 17(1-2):109–116, Aug. 1995.
62. T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using HMMs with dynamic features. In *Proc. ICASSP*, 1996.
63. T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Voice characteristics conversion for HMM-based speech synthesis system. In *Proc. ICASSP*, 1997.
64. P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(8):2280–2290, October 2012.
65. G. Galou. Synthetic voice forgery in the forensic context: a short tutorial. In *Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG)*, pages 1–3, September 2011.
66. W. Cai, A. Doshi, and R. Valle. Attacking speaker recognition with deep generative models. *arXiv preprint arXiv:1801.02384*, 2018.
67. T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda. A robust speaker verification system against imposture using an HMM-based speech synthesis system. In *Proc. Eurospeech*, 2001.
68. L.-W. Chen, W. Guo, and L.-R. Dai. Speaker verification against synthetic speech. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 309–312, 2010.
69. T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice-Hall, Inc., 2002.
70. Z. Wu, E.S. Chng, and H. Li. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In *Proc. Interspeech*, 2012.
71. A. Ogihara, H. Unno, and A. Shiozakai. Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 88(1):280–286, jan 2005.
72. P.L. De Leon, B. Stewart, and J. Yamagishi. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In *Proc. Interspeech 2012*, Portland, Oregon, USA, September 2012.
73. Y. Stylianou. Voice transformation: a survey. In *Proc. ICASSP*, pages 3585–3588, 2009.
74. B.L. Pellom and J.H.L. Hansen. An experimental study of speaker verification sensitivity to computer voice-altered imposters. In *Proc. ICASSP*, volume 2, pages 837–840, 1999.

75. S.H. Mohammadi and A. Kain. An overview of voice conversion systems. *Speech Communication*, 88:65 – 82, 2017.
76. M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proc. ICASSP*, pages 655–658, 1988.
77. L.M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28(3):211–226, 1999.
78. A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP*, volume 1, pages 285–288, 1998.
79. Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142, 1998.
80. T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2222–2235, 2007.
81. K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In *Proc. Interspeech*, 2014.
82. V. Popa, H. Silen, J. Nurminen, and M. Gabbouj. Local linear transformation for voice conversion. In *Proc. ICASSP*, pages 4517–4520. IEEE, 2012.
83. Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu. Voice conversion with smoothed GMM and MAP adaptation. In *Proc. EUROSPEECH*, pages 2413–2416, 2003.
84. H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen. A study of mutual information for GMM-based spectral conversion. In *Proc. Interspeech*, 2012.
85. E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj. Voice conversion using partial least squares regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):912–921, 2010.
86. N. Pilkington, H. Zen, and M. Gales. Gaussian process experts for voice conversion. In *Proc. Interspeech*, 2011.
87. D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose. One-to-many voice conversion based on tensor representation of speaker space. In *Proc. Interspeech*, pages 653–656, 2011.
88. H. Zen, Y. Nankaku, and K. Tokuda. Continuous stochastic feature mapping based on trajectory HMMs. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2):417–430, 2011.
89. Z. Wu, T. Kinnunen, E.S. Chng, and H. Li. Mixture of factor analyzers using priors from non-parallel speech for voice conversion. *IEEE Signal Processing Letters*, 19(12), 2012.
90. D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu. Statistical voice conversion based on noisy channel model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1784–1794, 2012.
91. P. Song, Y.Q. Bao, L. Zhao, and C.R. Zou. Voice conversion using support vector regression. *Electronics letters*, 47(18):1045–1046, 2011.
92. E. Helander, H. Silén, T. Virtanen, and M. Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE Trans. Audio, Speech and Language Processing*, 20(3):806–817, 2012.
93. Z. Wu, E.S. Chng, and H. Li. Conditional restricted boltzmann machine for voice conversion. In *the first IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2013.
94. M. Narendranath, H. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16(2):207–216, 1995.
95. S. Desai, E. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad. Voice conversion using artificial neural networks. In *Proc. ICASSP*, pages 3893–3896. IEEE, 2009.
96. Y. Saito, S. Takamichi, and H. Saruwatari. Voice conversion using input-to-output highway networks. *IEICE Transactions on Information and Systems*, E100.D(8):1925–1928, 2017.
97. T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion using RNN pre-trained by recurrent temporal restricted boltzmann machines. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):580–587, 2015.

98. L. Sun, S. Kang, K. Li, and H. Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *Proc. ICASSP*, pages 4869–4873, April 2015.
99. D. Sundermann and H. Ney. VTLN-based voice conversion. In *Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on*, pages 556–559. IEEE, 2003.
100. D. Erro, A. Moreno, and A. Bonafonte. Voice conversion based on weighted frequency warping. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):922–931, 2010.
101. D. Erro, E. Navas, and I. Hernaez. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(3):556–566, 2013.
102. C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y.u Tsao, and H.-M. Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. In *Proc. Interspeech 2017*, pages 3364–3368, 2017.
103. H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari. Voice conversion using sequence-to-sequence learning of context posterior probabilities. In *Proc. Interspeech 2017*, pages 1268–1272, 2017.
104. F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *Proc. ICASSP 2018*, 2018.
105. K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda. Statistical voice conversion with wavenet-based waveform generation. In *Proc. Interspeech*, pages 1138–1142, 2017.
106. B. Gillet and S. King. Transforming F0 contours. In *Proc. EUROSPEECH*, pages 101–104, 2003.
107. C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1109–1116, 2006.
108. E. Helander and J. Nurminen. A novel method for prosody prediction in voice conversion. In *Proc. ICASSP*, volume 4, pages IV–509. IEEE, 2007.
109. Z. Wu, T. Kinnunen, E.S. Chng, and H. Li. Text-independent F0 transformation with non-parallel data for voice conversion. In *Proc. Interspeech*, 2010.
110. D. Lolive, N. Barbot, and O. Boeffard. Pitch and duration transformation with non-parallel data. *Speech Prosody 2008*, pages 111–114, 2008.
111. T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi. The voice conversion challenge 2016. In *Proc. Interspeech*, pages 1632–1636, 2016.
112. M. Wester, Z. Wu, and J. Yamagishi. Analysis of the voice conversion challenge 2016 evaluation results. In *Proc. Interspeech*, pages 1637–1641, 2016.
113. P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. Voice forgery using ALISP: indexation in a client memory. In *Proc. ICASSP*, volume 1, pages 17–20. IEEE, 2005.
114. D. Matrouf, J.-F. Bonastre, and C. Fredouille. Effect of speech transformation on impostor acceptance. In *Proc. ICASSP*, volume 1, pages I–I. IEEE, 2006.
115. T. Kinnunen, Z. Wu, K.A. Lee, F. Sedlak, E.S. Chng, and H. Li. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *Proc. ICASSP*, pages 4401–4404. IEEE, 2012.
116. D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. Text-independent voice conversion based on unit selection. In *Proc. ICASSP*, volume 1, pages I–I, 2006.
117. Z. Wu, A. Larcher, K.A. Lee, E.S. Chng, T. Kinnunen, and H. Li. Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. In *Proc. Interspeech*, Lyon, France, 2013.
118. F. Alegre, R. Vipperla, N. Evans, and B. Fauve. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In *European Conference on Signal Processing (EUSIPCO), 2012 EURASIP Conference on*, 2012.

119. P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi. Detection of synthetic speech for the problem of imposture. In *Proc. ICASSP*, pages 4844–4847, Dallas, USA, 2011.
120. Z. Wu, T. Kinnunen, E.S. Chng, H. Li, and E. Ambikairajah. 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)*, pages 1–5. IEEE, 2012.
121. F. Alegre, R. Vipperla, and N. Evans. Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals. In *Proc. Interspeech*, 2012.
122. F. Alegre, A. Amehraye, and N. Evans. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proc. ICASSP*, 2013.
123. Z. Wu, X. Xiao, E.S. Chng, and H. Li. Synthetic speech detection using temporal modulation feature. In *Proc. ICASSP*, 2013.
124. F. Alegre, R. Vipperla, A. Amehraye, and N. Evans. A new speaker verification spoofing countermeasure based on local binary patterns. In *Proc. Interspeech*, Lyon, France, 2013.
125. Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Interspeech*, 2015.
126. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. Lee. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *INTERSPEECH*, 2017.
127. Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King. SAS: A speaker verification spoofing database containing diverse attacks. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.
128. Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi. ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *http://www.spoofingchallenge.org/asvSpoof.pdf*, 2014.
129. T.B. Patel and H.A. Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Proc. Interspeech*, 2015.
130. S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin. STC anti-spoofing systems for the ASVspoof 2015 challenge. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5475–5479, 2016.
131. N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu. Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge. In *Proc. Interspeech*, 2015.
132. X. Xiao, X. Tian, S. Du, H. Xu, E.S. Chng, and H. Li. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In *Proc. Interspeech*, 2015.
133. M.J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis. Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. In *Proc. Interspeech*, 2015.
134. Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanili, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado. Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):588–604, June 2017.
135. H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K.A. Lee, and J. Yamagishi. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 296–303, 2018.
136. M. Todisco, H. Delgado, and N. Evans. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Proc. Odyssey: the Speaker and Language Recognition Workshop*, pages 283–290, Bilbao, Spain, June 21-24 2016.
137. M. Todisco, H. Delgado, and N. Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516 – 535, 2017.

138. G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proc. Interspeech*, pages 82–86, 2017.
139. Z. Ji, Z.Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao. Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017. In *Proc. Interspeech*, pages 87–91, 2017.
140. L. Li, Y. Chen, D. Wang, and T.F. Zheng. A study on replay attack and anti-spoofing for automatic speaker verification. In *Proc. Interspeech*, pages 92–96, 2017.
141. H.A. Patil, M.R. Kamble, T.B. Patel, and M.H. Soni. Novel variable length teager energy separation based instantaneous frequency features for replay detection. In *Proc. Interspeech*, pages 12–16, 2017.
142. Z. Chen, Z. Xie, W. Zhang, and X. Xu. ResNet and model fusion for automatic spoofing detection. In *Proc. Interspeech*, pages 102–106, 2017.
143. Z. Wu, S. Gao, E.S. Cling, and H. Li. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–5. IEEE, 2014.
144. Q. Li. An auditory-based transform for audio signal processing. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 181–184. IEEE, 2009.
145. S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, Aug 1980.
146. M. Sahidullah, T. Kinnunen, and C. Hanilci. A comparison of features for synthetic speech detection. In *Proc. Interspeech*, pages 2087–2091. ISCA, 2015.
147. J. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustic Society of America*, 89(1):425–434, 1991.
148. M.J. Alam and P. Kenny. Spoofing detection employing infinite impulse response - constant Q transform-based feature representations. In *Proc. European Signal Processing Conference (EUSIPCO)*, 2017.
149. P. Cancela, M. Rocamora, and E. López. An efficient multi-resolution spectral transform for music analysis. In *Proc. International Society for Music Information Retrieval Conference*, pages 309–314, 2009.
150. Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.
151. I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. MIT Press, Cambridge, 2016.
152. Y. Tian, M. Cai, L. He, and J. Liu. Investigation of bottleneck features and multilingual deep neural networks for speaker verification. In *Proc. Interspeech*, pages 1151–1155, 2015.
153. F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675, Oct 2015.
154. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012.
155. M.J. Alam, P. Kenny, V. Gupta, and T. Stafylakis. Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks. In *Proc. Odyssey: the Speaker and Language Recognition Workshop*, pages 270–276, Bilbao, Spain, June 21-24 2016.
156. Y. Qian, N. Chen, and K. Yu. Deep features for automatic spoofing detection. *Speech Communication*, 85:43–52, 2016.
157. H. Yu, Z. H. Tan, Y. Zhang, Z. Ma, and J. Guo. DNN filter bank cepstral coefficients for spoofing detection. *IEEE Access*, 5:4779–4787, 2017.
158. K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li. Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):632–643, June 2017.
159. J. Andén and S. Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

160. S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65:1331–1398, 2012.
161. M. Pal, D. Paul, and G. Saha. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language*, 48:31–50, 2018.
162. K. Laskowski, M. Heldner, and J. Edlund. The fundamental frequency variation spectrum. In *Proceedings of FONETIK*, volume 2008, pages 29–32, 2008.
163. I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas. Synthetic speech detection using phase information. *Speech Communication*, 81:30–41, 2016.
164. L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami. Spoofing speech detection using modified relative phase information. *IEEE Journal of selected topics in signal processing*, 11(4):660–670, 2017.
165. S. Chakroborty and G. Saha. Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. *International Journal of Signal Processing*, 5(1):11–19, 2009.
166. X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.
167. A. R. Goncalves, R. P. V. Violato, P. Korshunov, S. Marcel, and F. O. Simoes. On the generalization of fused systems in voice presentation attack detection. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, Sept 2017.
168. D. Paul, M. Pal, and G. Saha. Novel speech features for improved detection of spoofing attacks. In *Proc. Annual IEEE India Conference (INDICON)*, 2016.
169. N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 19(4):788–798, May 2011.
170. E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel. Introducing i-vectors for joint anti-spoofing and speaker verification. In *Proc. Interspeech*, 2014.
171. A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel. Joint speaker verification and anti-spoofing in the i-vector space. *IEEE Trans. Information Forensics and Security*, 10(4):821–832, 2015.
172. C. Hanilçi. Data selection for i-vector based automatic speaker verification anti-spoofing. *Digital Signal Processing*, 72:171–180, 2018.
173. X. Tian, Z. Wu, X. Xiao, E.S. Chng, and H. Li. Spoofing detection from a feature representation perspective. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2119–2123, March 2016.
174. H. Yu, Z. H. Tan, Z. Ma, R. Martin, and J. Guo. Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–12, 2018.
175. H. Dinkel, N. Chen, Y. Qian, and K. Yu. End-to-end spoofing detection with raw waveform cldnns. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4860–4864, March 2017.
176. T.N. Sainath, R.J. Weiss, A. Senior, K.W. Wilson, and O. Vinyals. Learning the speech front-end with raw waveform CLDNNs. In *Proc. Interspeech*, 2015.
177. C. Zhang, C. Yu, and J. H. L. Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):684–694, June 2017.
178. H. Muckenhirn, M. Magimai-Doss, and S. Marcel. End-to-end convolutional neural network-based voice presentation attack detection. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 335–341, Oct 2017.
179. S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*, pages 183–195. IEEE, 2017.
180. S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Proc. Interspeech*, 2015.

181. S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui. Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector. In *ODYSSEY*, 2016.
182. M. Sahidullah, D.A.L. Thomsen, R.G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen. Robust voice liveness detection and speaker verification using throat microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):44–56, 2018.
183. G.W. Elko, J. Meyer, S. Backer, and J. Peissig. Electronic pop protection for microphones. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 46–49. IEEE, 2007.
184. L. Zhang, S. Tan, J. Yang, and Y. Chen. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1080–1091. ACM, 2016.
185. L. Zhang, S. Tan, and J. Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 57–71. ACM, 2017.
186. C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov. Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. *Speech Communication*, 85:83 – 97, 2016.
187. H. Yu, A.K. Sarkar, D.A.L. Thomsen, Z.-H. Tan, Z. Ma, and J. Guo. Effect of multi-condition training and speech enhancement methods on spoofing detection. In *Proc. International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 7 2016.
188. X. Tian, Z. Wu, X. Xiao, E.S. Chng, and H. Li. An investigation of spoofing speech detection under additive noise and reverberant conditions. In *Proc. Interspeech*, 2016.
189. H. Delgado, M. Todisco, N. Evans, M. Sahidullah, W.M. Liu, F. Alegre, T. Kinnunen, and B. Fauve. Impact of bandwidth and channel variation on presentation attack detection for speaker verification. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2017.
190. Y. Qian, N. Chen, H. Dinkel, and Z. Wu. Deep feature engineering for noise robust spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1942–1955, 2017.
191. P. Korshunov and S. Marcel. Cross-database evaluation of audio-based spoofing detection systems. In *Proc. Interspeech*, 2016.
192. D. Paul, M. Sahidullah, and G. Saha. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2047–2051. IEEE, 2017.
193. J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen. Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. In *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2018.
194. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
195. F. Kreuk, Y. Adi, M. Cisse, and J. Keshet. Fooling end-to-end speaker verification by adversarial examples. *arXiv preprint arXiv:1801.03339*, 2018.
196. M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan. Integrated spoofing countermeasures and automatic speaker verification: an evaluation on ASVspoof 2015. In *Proc. Interspeech*, 2016.
197. H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel. Long-term spectral statistics for voice presentation attack detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(11):2098–2111, 2017.
198. A. Sarkar, M. Sahidullah, Z.-H. Tan, and T. Kinnunen. Improving speaker verification performance in presence of spoofing attacks using out-of-domain spoofed data. In *Proc. Interspeech*, 2017.

199. T. Kinnunen, K.A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D.A. Reynolds. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. In *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2018.

200. M. Todisco, H. Delgado, K.A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi. Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion. In *Proc. Interspeech*, 2018.

201. Z. Wu, P.L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and Y. Yamagishi. Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):768–783, 2016.