An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks

Artur Janicki, Federico Alegre and Nicholas Evans

Abstract

This article analyses the threat of replay spoofing or presentation attacks in the context of automatic speaker verification (ASV). As relatively high-technology attacks, speech synthesis and voice conversion, which have thus far received far greater attention in the literature, are probably beyond the means of the average fraudster. The implementation of replay attacks, in contrast, requires no specific expertise nor sophisticated equipment. Replay attacks are thus likely to be the most prolific in practice, even if their impact is relatively under-researched. This paper compares the threat of replay attacks to those of speech synthesis and voice conversion. The comparison is performed using strictly controlled protocols and with six different ASV systems including a state-of-the-art iVector system with probabilistic linear discriminant analysis. Experiments show that low-effort replay attacks present at least a comparable threat to speech synthesis and voice conversion. The paper also describes and assesses two replay attack countermeasures. A relatively new approach based on the local binary pattern (LBP) analysis of speech spectrograms is shown to outperform a competing approach based on the detection of far-field recordings.

Index Terms

speaker verification, spoofing, replay, countermeasures, local binary patterns.

A. Janicki is with the Institute of Telecomunications, Warsaw University of Technology, Warsaw, Poland, e-mail: A.Janicki@tele.pw.edu.pl.

F. Alegre and N. Evans are with EURECOM, Sophia Antipolis, France, e-mail: {alegre,evans}@eurecom.fr.

A. Janicki was supported by the European Union in the framework of the European Social Fund through the Warsaw University of Technology Development Programme.

F. Alegre and N. Evans were supported by the TABULA RASA project funded under the 7th Framework Programme of the European Union (grant agreement number 257289).

Manuscript received MMMM DD, YYYY; revised MMMM DD, YYYY.

I. INTRODUCTION

Spoofing refers to the presentation of a falsified or manipulated sample to the sensor of a biometric system in order to provoke a high score and thus illegitimate verification. In recent years the automatic speaker verification (ASV) community has started to investigate spoofing and countermeasures actively [1], [2]. A growing body of independent work has demonstrated the vulnerability of ASV systems to spoofing through replayed speech [3], [4], impersonation [5], [6], voice conversion [7], [8], speech synthesis [9], [10] and attacks with non-speech, artificial, tone-like signals [11], [12].

Common to the bulk of previous work is a focus on attacks which require either specific expertise, e.g. impersonation, or high-level technology, e.g. speech synthesis and voice conversion. Only replay attacks can be performed with ease, requiring neither specialist expertise nor sophisticated equipment. They are easily implemented with discreet, high-quality audio equipment, now available to the masses. Accordingly, it is reasonable to assume that replay attacks represent a tangible threat and that they will be the most prolific in practice.

Only few studies have addressed replay. The work in [3] assessed the vulnerabilities of a hidden Markov model (HMM) based, text-dependent ASV system using concatenated digits. While results showed that replay attacks are highly effective, experiments were conducted with data collected from only two speakers. The work in [4] investigated replay using recordings collected with close-talk or far-field microphones and then replayed over an analogue or digital telephony channel. The work was conducted with data collected from five speakers and demonstrated the vulnerability of a joint factor analysis (JFA) ASV system; the false acceptance rate (FAR) at the equal error rate (EER) threshold increased from 1% to almost 70%. The authors in [13] investigated a text-dependent ASV system which was subjected to speech replayed using a laptop computer. Using the large, standard and publicly available RSR2015 corpus, this work showed that the EER for an HMM system increased from approximately 4% to more than 20%.

Missing from the literature, however, is a reliable comparative assessment of replay attacks to speech synthesis and voice conversion using large, standard databases and using a representative range of replay scenarios. Such a study is needed in order to help prioritise future work to develop countermeasures for the protection of ASV from spoofing. While well-intentioned, the work concerning speech synthesis and voice conversion attacks (including that of the authors), may have over-exaggerated the threat given that only few people have the necessary capabilities to implement them. Meanwhile, without appropriate countermeasures, ASV systems may remain vulnerable to replay attacks which are implemented much

Attack	Naïve impostor	Replay	Voice conversion	Speech synthesis
Input	impostor speech	target speech	impostor speech	text
Effort	zero	low	medium-high	high
Effectiveness	low	**(?)**	medium-high	high

TABLE I: Comparison of four different attacks in terms of speech used, required effort and effectiveness.

more easily. This paper accordingly aims to assess ASV vulnerabilities to replay attacks using the same ASV systems and corpora used in previous assessments involving speech synthesis and voice conversion. In addition, the paper investigates the effectiveness of new countermeasures which aim to distinguish between genuine and replayed speech.

The paper is organised as follows. Section II describes speech synthesis and voice conversion spoofing attacks with a comparison to replay attacks. Specific implementations used for the work reported here are also described. Section III presents previous and ongoing work to develop countermeasures against replay attacks, including the authors' work using the local binary pattern analysis of speech spectrograms. A common experimental framework for the assessment of both vulnerabilities and countermeasures is presented in Section IV. Results are presented in Section V. Priorities for future work are discussed in Section VI followed by conclusions in Section VII.

II. Spoofing speaker verification systems

This section describes speech synthesis, voice conversion and replay spoofing attacks and specific implementations. In general, a spoofed speech signal $s(t)$ is generated from the speech signal of a target speaker $x(t)$. Whereas the input to a speech synthesis system is a text string, that to voice conversion originates from a third speech signal of a source speaker (spoofer) $y(t)$.

A. Speech synthesis

There is a large variety of speech synthesis algorithms, such as formant [14], diphone [15], unit-selection [16] and statistical parametric [17] based approaches, in addition to more recent deep-neural network architectures. Whatever the approach, the aim is to generate intelligible, natural speech for a given text string c. In the context of spoofing, a synthetic speech signal is generated according to:

$$s(t) = g_{x(t)}(c), \tag{1}$$

where $g_{x(t)}$ denotes a text-to-speech mapping generated by a synthesis system with speech units or acoustic models extracted or learned from the speech signal of a target speaker $x(t)$. While unit-selection approaches generally require large amounts of speaker-specific data to learn the mapping function $g_{x(t)}$, statistical parametric approaches can synthesize convincing speech signals with the adaptation of well-trained models using relatively small quantities of speaker-specific data [18].

Our approach to statistical parametric speech synthesis uses hidden Markov models following the approach described in [17]. The specific implementation uses the HMM-based Speech Synthesis System (HTS)[1] where speech signals are parametrised by STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) features, Mel-cepstrum coefficients and the logarithm of the fundamental frequency (log F_0) with their delta and delta-delta coefficients. Acoustic spectral characteristics and duration probabilities are modelled using multispace distribution hidden semi-Markov models (MSD-HSMM) [19]. Speaker dependent excitation, spectral and duration models are adapted from corresponding independent models according to a speaker adaptation strategy referred to as constrained structural maximum a posteriori linear regression (CSMAPLR) [20]. Finally, time domain signals are synthesised using a vocoder based on Mel-logarithmic spectrum approximation (MLSA) filters. They correspond to STRAIGHT Mel-cepstral coefficients and are driven by a mixed excitation signal and waveforms reconstructed using the pitch synchronous overlap add (PSOLA) method.

B. Voice conversion

Voice conversion has been used to explore ASV spoofing since the late 90s [8]. One successful approaches involves so-called Gaussian-dependent filtering [21]. Here, the spoofing signal $s(t)$ (or $S(f)$ in the spectral domain) is generated by filtering at the frame level the speech signal of the source or spoofer $y(t)$. In the spectral domain it can be represented as follows:

$$S(f) = \frac{|H_x(f)|}{|H_y(f)|} Y(f) \tag{2}$$

where $H_x(f)$ and $H_y(f)$ are the vocal tract transfer functions of the targeted speaker and the spoofer respectively. $Y(f)$ is the spoofer's speech signal in the spectral domain whereas $S(f)$ denotes the result after voice conversion. A time-domain, converted signal is recovered as follows:

$$s(t) = IFFT(\frac{|H_x(f)|}{|H_y(f)|}) * y(t) \tag{3}$$

[1]http://hts.sp.nitech.ac.jp/

where * denotes convolution. As such, $y(t)$ is mapped or converted towards the target in a spectral-envelope sense, and is sufficient to overcome most ASV systems [21], [22].

$H_x(f)$ is determined from a set of two Gaussian mixture models (GMMs). The first, denoted as the automatic speaker recognition (asr) model in the original work, is related to ASV feature space and utilised for the calculation of a posteriori probabilities. The second, denoted as the filtering (fil) model, is a tied model of linear predictive cepstral coding (LPCC) coefficients from which $H_x(f)$ is derived. LPCC filter parameters are obtained according to:

$$x_{fil} = \sum_{i=1}^{M} p(g_{asr}^i | y_{asr}) \mu_{fil}^i \tag{4}$$

where $p(g_{asr}^i | y_{asr})$ is the a posteriori probability of Gaussian component g_{asr}^i given the frame y_{asr} and μ_{fil}^i is the mean of component g_{fil}^i which is tied to g_{asr}^i. $H_x(f)$ is estimated from x_{fil} using an LPCC-to-LPC transformation and a time-domain signal is synthesised from converted frames with a standard overlap-add technique. Full details can be found in [21], [22], [23].

C. Replay

Replay attacks are an example of low-effort spoofing; they require simply the replaying of a previously captured speech signal. In the absence of suitable countermeasures and considering the widespread availability of consumer devices with high-quality sound systems, replay attacks can typically be realised with ease. Furthermore, used either directly, or through the cutting and pasting of short speech intervals, replayed speech has potential to overcome both text-dependent and text-independent ASV systems. Even though the processes of recording and replaying introduce additive acoustic and convolutive channel and transducer noise, these effects can be attenuated by noise and other intersession (channel) variability compensation techniques common to most modern speaker recognition systems. These factors point towards the tangible threat posed by replay attacks.

Ignoring ambient noise in the acoustic environment (which is in any case not specific to the replay spoofing scenario), replayed speech can be represented as:

$$s(t) = x(t) * h(t), \tag{5}$$

where * denotes convolution. The composite replay effects denoted by $h(t)$ include the impulse responses of replay hardware and the replay environment. It is composed by:

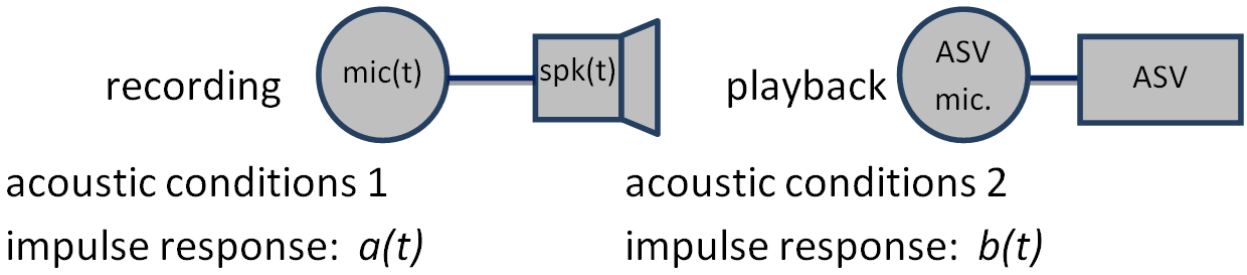

Fig. 1: A schematic diagram of the assumed replay attack configuration.

$$h(t) = mic(t) * a(t) * spk(t) * b(t), \tag{6}$$

where $mic(t)$ and $spk(t)$ are impulse responses of the microphone and the speaker, respectively, and where $a(t)$ and $b(t)$ are the respective impulse responses of the recording and replay environments. This scenario is illustrated in Fig. 1. Replay attacks may thus be emulated through the convolution of a source signal $s(t)$ with typical impulse responses representing the different replay components.

Since an inordinately large number of different impulse responses would be needed to emulate representative replay attacks using Equation 5, a simplification can be applied to reduce the number of impulse responses. Since typical loudspeaker responses deviate more so from a regular impulse than typical microphone responses, and thus dominate Equation 6, $mic(t)$ can be safely ignored. Furthermore, since $a(t)$ and $b(t)$ are similar in nature, a reasonable study can be achieved by considering only one. Thus:

$$h(t) \simeq spk(t) * b(t) \tag{7}$$

Equation 7 represents a worst-case scenario where the spoofer obtains a high quality recording of the target speaker's voice, with no recording artefacts. In this case, only replay components are considered and are the only means by which a replay attack can be detected.

D. Qualitative comparison

Replay, voice conversion and speech synthesis spoofing are forms of *concerted-effort* impostor attacks, as opposed to the naïve or *zero-effort* impostor attacks normally used to assess ASV system performance. A qualitative comparison of all four is illustrated in Table I, ordered by the level of effort or expertise needed to implement each attack [2].

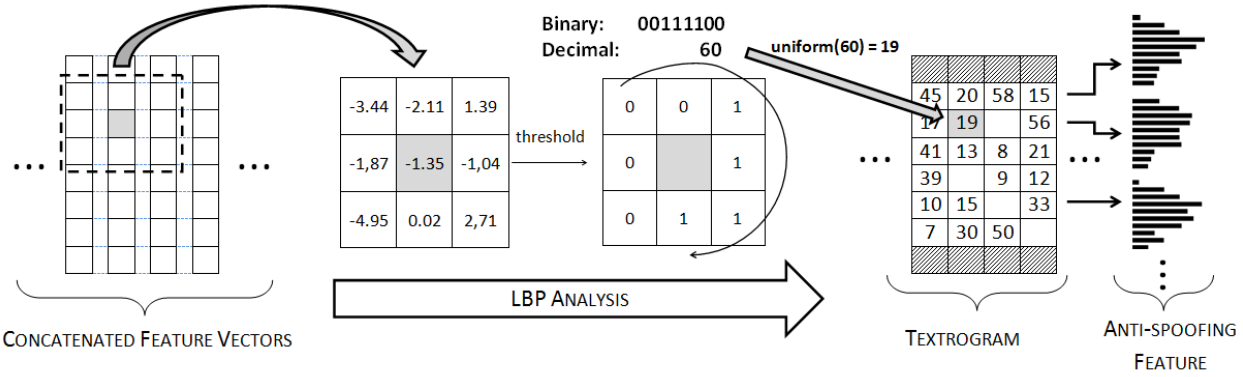

Fig. 2: Schematic diagram of LBP-based feature extraction.

Compared to naïve impostor attacks, replay attacks require slightly increased effort; they require recording and replaying. Voice conversion and speech synthesis attacks require specialised, often complex algorithms, in addition to any recording hardware to collect, analyse and parametrise the target and any other auxiliary speech data. They belong to a class of higher-effort spoofing attacks. While voice conversion is based upon the conversion of one speech signal to another, speech synthesis converts a text string to a speech signal, which requires a comparatively higher level of effort or expertise.

One may reasonably suppose that the effectiveness of each attack is correlated with the level of effort involved in their implementation; the higher the effort, the more effective the algorithm and hence the greater the impact on ASV performance. Replay attacks are then assumed to pose only a low threat. However, the work in [24], [13] suggests the contrary, showing that replay attacks pose a significant threat, being effective in overcoming an ASV system while being the most easily implemented spoofing attack. This paper investigates these contradictory findings and compares objectively and quantitatively the threat of replay spoofing to those of speech synthesis and voice conversion.

III. Replay Countermeasures

Attention now turns to the detection of replay spoofing attacks with dedicated countermeasures. Given that only little work has investigated ASV vulnerabilities to such attacks, it is hardly surprising that work to develop countermeasures is similarly limited. This section briefly reviews that past work and then describes two particular replay countermeasures which are explored further in this paper.

A. General approaches

One obvious approach to replay detection involves challenge-response systems which require the speaker to utter a prompted phrase [25]. Challenge-response mechanisms are a form of passive countermeasure. While having potential in preventing some forms of replay attack for some ASV systems, challenge-response countermeasures are not without impacts on usability which may render them undesirable for other ASV systems.

Active countermeasures have also been proposed. One such approach involves the storing of previous access attempts and their comparison to new attempts [26]. New access attempts which are deemed too close to previous attempts are rejected. A somewhat similar technique is proposed in [13], where the authors compare spectral bitmaps between access trials and previously stored recordings in a text-dependent ASV scenario.

Other, more generally applicable methods not restricted to any particular ASV scenario, are based on the detection of unexpected channel artefacts indicative of recording and replaying. Two such algorithms are reported in [27] for which the EER for a baseline Gaussian mixture model (GMM) system with a universal background model (UBM) was shown to decrease from 40% to 10% with active countermeasures. Channel detection is the basis of the first approach investigated further in this paper.

B. Far-field channel detection

Many scenarios in which user authentication is performed by ASV involve so-called close-talk speech, i.e. situations where speech is collected from an in-situ or closely positioned microphone. Examples include telephony and logical access scenarios or critical infrastructure protection and physical access scenarios. In contrast, since they are likely to be collected surreptitiously or at-distance, replay recordings may exhibit far-field channel effects, effects which can be measured and consequently used to detect replayed speech.

This idea was first investigated in [28]. The work compares close-talk and far-field speech signals parametrised according to 12 channel-sensitive features:

- spectral ratio – sub-band energy ratio from 0-2 kHz and from 2-4 kHz;
- low frequency ratio – sub-band energy ratio from 100-300 Hz and from 300-500 Hz, calculated using speech frames only;
- modulation index, and
- nine sub-band modulation indices – see [28] for precise sub-band bandwidths.

The spectral ratio reflects the level of spectrum flattening or noise and reverberation introduced by far-field recording. The low frequency ratio reflects the (potentially heightened) level of high-pass filtering, an artefact typical of speech signals produced by small loudspeakers. The total and sub-band modulation indices reflect the level of additive and, specifically, coloured noise; higher levels of noise present in replay recordings result in lower than average modulation indices. Experiments were performed with a support vector machine (SVM) classifier and showed that far-field recordings could be detected with 90% accuracy.

C. Local binary patterns

The approach to replay detection proposed in [24] is based on the local binary pattern (LBP) analysis of speech spectrograms. Inspired by the original application to image texture analysis [29], the idea was introduced as an ASV spoofing countermeasure in [30]. As illustrated in Fig. 2, LBP analysis is applied to a Mel-scaled cepstrogram with appended dynamic features. Modifications made through spoofing are assumed to disturb the natural 'texture' of genuine speech, attributes which are readily detected with LBP.

The standard LBP operator is a non-parametric 3x3 kernel which assigns a binary code to each pixel in an image according to the comparison of its intensity value to that of its eight surrounding pixels [29]. A binary value of 1 is assigned when the intensity of neighbouring pixels (here feature components) is higher, whereas a value of 0 is assigned when neighbouring pixels are of lower or equal intensity. Each pixel is thus assigned one of $2^8 = 256$ binary patterns.

LBPs can be determined for each pixel in a Mel-scaled cepstrogram thus resulting in a new matrix of reduced dynamic range, here referred to as a *textrogram*. The textrogram captures short-time feature motion beyond that in conventional dynamic parametrisations. Normalised histograms of pixel values constructed for each row of the textrogram are stacked vertically to obtain the anti-spoofing feature vector in the same manner as GMM mean-vectors are stacked to form supervectors. In the work reported here, LBP-based features are calculated for the both the enrolment and test data. The two resulting feature vectors are compared using histogram intersection and the resulting score is thresholded to decide if the test signal is genuine speech or a spoofing attack. Experimental results presented in [30] showed that the LBP-based textrogram analysis is effective in detecting a range of spoofed speech signals, including artificial signals and speech synthesis (EERs of below 1%) though less effective in the case of voice conversion (EER in the order of 7%).

IV. EXPERIMENTAL SETUP

The comparison of replay, speech synthesis and voice conversion spoofing attacks necessitates controlled experiments with a large, standard corpus, a variety of ASV systems and judicious experimental protocols. They are described here.

A. Methodology

This work aims to compare in a meaningful manner the threat of replay, speech synthesis and voice conversion spoofing attacks. The goal places significant constraints on the choice of dataset, limiting the scope to those for which speech synthesis and voice conversion assessments have already been reported. Large, standard datasets are also preferred.

The work reported here is performed on a subset of the public corpora released in the context of the speaker recognition evaluation (SRE) campaigns administered by the National Institute for Standards and Technology (NIST) [31]. While the use of standard databases enables the comparison of results to others' work, it also necessitates artificial replay emulation. While not ideal, this approach is preferred since it enables the comparison of results for replay, speech synthesis and voice conversion using otherwise identical protocols (and same source data). It is stressed, however, that the work uses impulse responses measured using real playback hardware and real acoustic environments.

B. ASV systems

Since previous work in ASV anti-spoofing has shown different vulnerabilities for different ASV systems, this work is similarly performed with a range of representative technologies, from the standard to the state-of-the-art. They include: (i) a standard Gaussian mixture model with universal background model (GMM-UBM) system; (ii) a GMM supervector linear kernel (GSL) system; (iii) a GSL system with nuissance atribute projection (NAP) [32]; (iv) a GSL system with factor analysis (FA) [33]; (v) a GMM-UBM system with factor analysis; (vi) an iVector system [34] with probabilistic linear discriminant analysis (PLDA) [35] and length normalisation [36] (referred to from here on as IV-PLDA). Experiments were performed with and without score normalisation. Symmetric normalisation (S-norm) [37] was applied to the IV-PLDA system while test normalisation (T-norm) [38] was used for the others.

All ASV systems are implemented with the same LIA-SpkDet toolkit [39] and the ALIZE library [40] and stem from the work in [33]. They use a common UBM with 1024 Gaussian components and a common feature parametrisation: linear frequency cepstral coefficients (LFCCs), their deltas and delta

energy. A speech activity detector is also common to each system. It fits a 3-component GMM to the log-energy distribution and adjusts the speech/non-speech threshold according to the GMM parameters [41]. This approach has been used successfully in many independent studies, e.g. [43].

C. Datasets, protocols and metrics

Experiments are performed on the male subsets of the 2004, 2005, 2006 and 2008 NIST SRE datasets, which are from here on referred to as NIST'0x. The NIST'04 and NIST'08 datasets are used for UBM training. The NIST'05 dataset is used for development, whereas the NIST'06 dataset is used for evaluation; only results for the latter are reported here. Due to the significant amount of data necessary to estimate the total variability matrix T used with the IV-PLDA system, the NIST'06 dataset is added to the pool of background data for development whereas the NIST'05 dataset was used for evaluation. T is thus learned using approximately 11,000 utterances from 900 speakers, while independence between development and evaluation datasets is always respected.

All experiments relate to the 8conv4w-1conv4w condition where one conversation corresponds to an average of 2.5 minutes of speech (one side of a 5 minute conversation). In all cases, however, only one of the eight, randomly selected training conversations is used for enrolment. Results presented in this paper are thus comparable only to those in the literature (not work related to spoofing) for the 1conv4w-1conv4w condition. Standard NIST protocols dictate in the order of 1,000 true client tests and 10,000 impostor tests for development and evaluation datasets. To assess spoofing, impostor tests are replaced with spoofed versions of the original utterance, whereas genuine client trials are unchanged. This setup conforms to the general protocols outlined in [2] and the broader evaluation methodology described in [44].

Given the consideration of spoofing, and without any standard operating criteria under such a scenario, the equal error rate (EER) metric is preferred to the minimum detection cost function (minDCF). Also reported is the spoofing false acceptance rate (SFAR, [45]) for a false rejection rate (FRR) fixed to the baseline EER.

D. Spoofing emulation

Spoofing attacks are generally assumed to be performed at the microphone level. In the case of voice conversion and speech synthesis, the practical scenario would then involve the recording of suitable data for the training or adaptation of conversion or synthesis systems, the fabrication of a spoofed utterance and then its presentation to the microphone of the ASV system. None of the past work (including the

authors') follows this process, preferring instead to emulate spoofing attacks by intervening after the microphone, immediately prior to feature extraction. This approach is entirely justified in the case of linear transducers and channels, and mainly telephony scenarios in which spoofing can be applied while bypassing the microphone [46].

Voice conversion spoofing attacks are emulated with the approach described in Section II-B. A worst-case scenario is considered; conversion is performed with full prior knowledge of the ASV system, i.e. voice conversion is performed with exactly the same front-end processing as that used for ASV. $H_y(f)$ and filter $H_x(f)$ use 19 LPCC and LPC coefficients respectively. Voice conversion is applied to the original impostor utterances which are converted towards the genuine speaker for any given trial.

Speech synthesis attacks are emulated according to the approach described in Section II-A using the voice cloning toolkit[2] with a default configuration and standard speaker-independent models trained on the EMIME corpus [47]. Adaptation data for each target speaker comprises three utterances (with transcriptions). For any given trial, speech synthesis spoofing attacks are generated using arbitrary text, thereby producing a spoofed utterance of duration close to that of the average test utterance.

Replay attacks are emulated according to the approach described in Section II-C using a random mix of three different loudspeakers impulse responses $spk(t)$ and three different replay environments $b(t)$. Speaker impulse responses $spk(t)$ are obtained from [48] and correspond to a low-quality smartphone speaker, a medium-quality tablet speaker and a high-quality stand-alone speaker. The impulse response and frequency responses of each are illustrated in Fig. 3. There are significant differences in the frequency responses which show in particular the high-pass functions of the smaller (and perhaps lower quality) devices. The third impulse response is comparatively flat, though with greater high-frequency attenuation. The first two impulse responses are, however, comparatively short, whereas the third extends to 2ms. The first two replay environment impulse responses $b(t)$ (not illustrated here) are obtained from [49] and correspond to an enclosed medium-sized office and an open corridor. The third impulse response simulates an anechoic chamber with a flat frequency response.

E. Countermeasures

The far-field channel detection countermeasure is implemented according to the algorithm originally proposed in [28] and described in Section III-B. The modulation indices are calculated frame-wise from the speech signal envelope which is approximated by the absolute value of the signal after down-sampling

[2]http://homepages.inf.ed.ac.uk/jyamagis/software/page37/page37.html

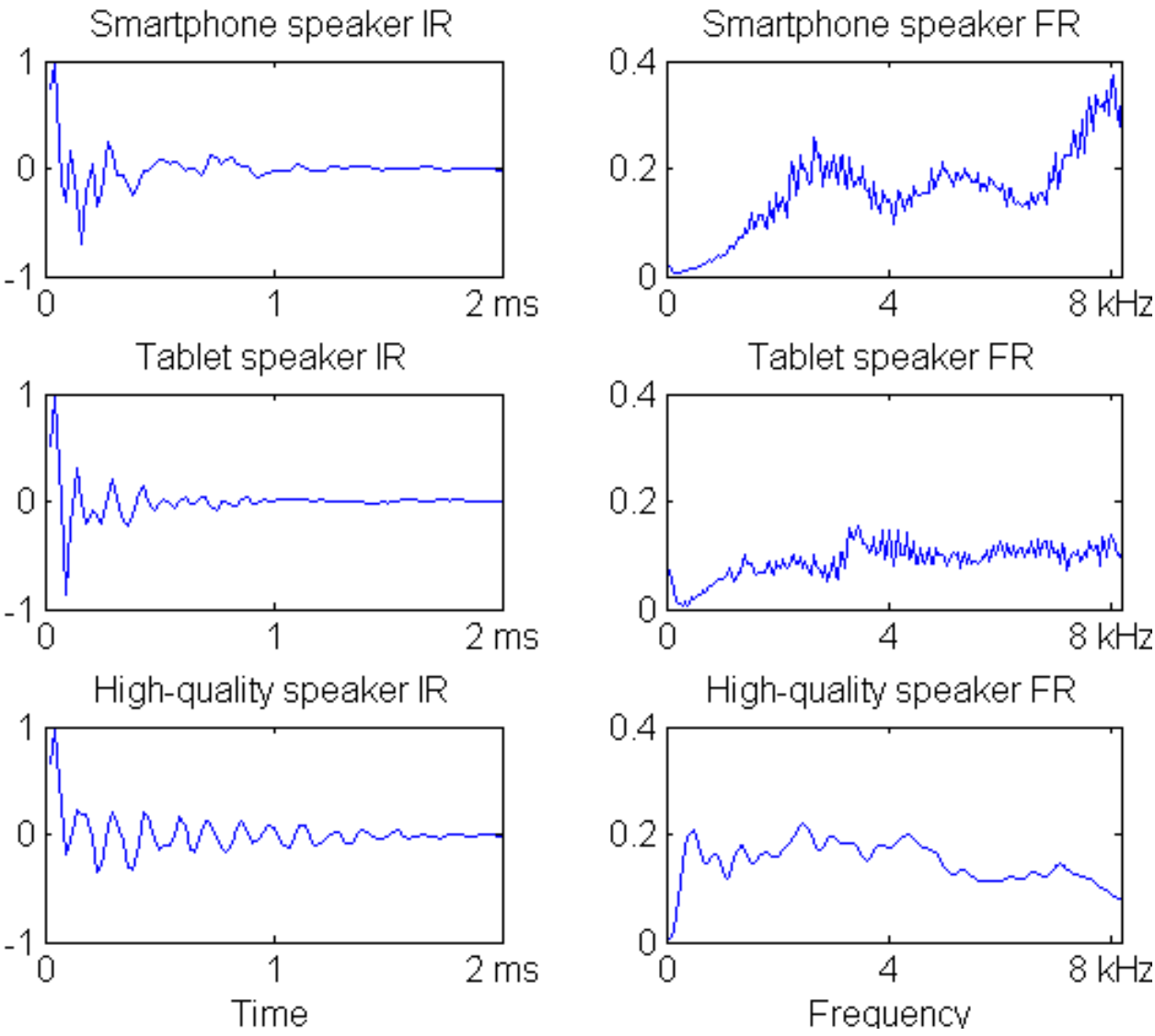

Fig. 3: Impulse (left) and frequency (right) responses for three different speakers.

to 60 Hz. The final modulation index is calculated by averaging over speech frames whose modulation index is above 0.75.

The LBP countermeasure is implemented using the toolkit provided by The University of Oulu[3]. Normalised acoustic features used for LBP analysis are composed of 51 coefficients: 16 LFCCs and energy plus their corresponding delta and delta-delta coefficients. Analysis is applied only to speech

[3]http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

frames and using only the 58 so-called uniform LBPs[4] as originally described in [29] and for speech processing in [30]. LBP histograms are created for all but the first and last rows, i.e., for $51 - 2 = 49$ rows. Non-uniform LBPs are ignored thereby resulting in feature vectors of $58 \times 49 = 2842$ dimensions.

Both countermeasure algorithms are trained using a random subset of 1000 utterances from the NIST'05 dataset which are treated as described in Section IV-C in order to generate suitable training data with various acoustic conditions. Room and loudspeaker impulse responses (a lecture room, a staircase and a meeting room) different to those used for ASV experiments aims to minimise countermeasure over-fitting.

A Bayesian network classifier [50] is learned to differentiate genuine data from spoofed data in the case of the far-field channel detection countermeasure. In contrast, the AdaBoost M1 meta classifier [51] is used for the LBP countermeasure. These classifiers returned the best results in each case for an area-under-the-ROC metric.

For evaluation, countermeasures are used independently from ASV systems, similarly to the protocol used, e.g., in the ASVspoof 2015 evaluation [52] (which does not include replay attacks). There are 1352 genuine trials and 8112 replay attacks, using various replay hardware and acoustic environments. In a second set of experiments, countermeasures are integrated with ASV as in [30]. In this case, the countermeasure threshold is set heuristically to minimise the EER of the ASV system. Trials classified as spoofs are assigned an arbitrarily low score and are thereby rejected automatically.

V. Results

Attention now turns to the assessment of ASV vulnerabilities. With a strictly controlled protocol, the potential impact of replay attacks is compared to that of speech synthesis and voice conversion. It is stressed that the experiments do not and cannot evaluate every possible replay scenario in exhaustive fashion; the aim is simply to gauge the relative threat and the potential to detect attacks with dedicated countermeasures.

A. Replay spoofing

Fig. 4 shows the effect of replay attacks on the score distributions for the IV-PLDA system. Replay attacks correspond to an emulated high-quality speaker and office environment. Three distributions are illustrated: (i) the zero-effort impostor distribution; (ii) the genuine client distribution, and (iii) the

[4]The subset of LBPs which contain at most two bitwise transitions from 0 to 1 or 1 to 0 when the bit pattern is traversed in circular fashion

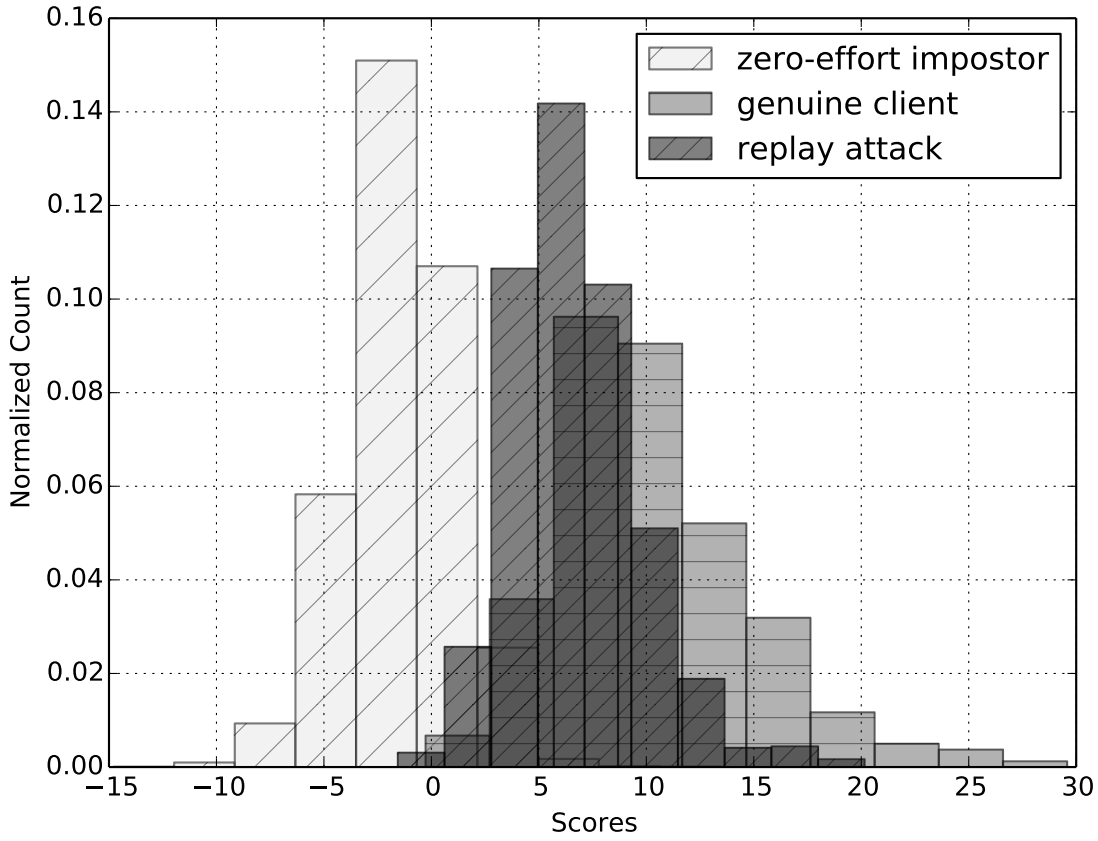

Fig. 4: Score distribution for the IV-PLDA system for replay attacks using an emulated high-quality speaker and an office environment.

distribution obtained when all zero-effort impostor trials are replaced with replay attacks. While the zero-effort impostor and genuine client distributions are well separated, the latter overlaps significantly with the score distribution for replay attacks. Increased overlap between these distribution will degrade ASV performance.

Fig. 5 illustrates a detection error trade-off (DET) plot[5] for the same experimental setup. The lower-most, solid profile illustrates the performance of the baseline ASV system. The upper-most, dashed profile illustrates performance when zero-effort impostors are replaced with replay attacks. The difference

[5]The DET plots were produced with the TABULA RASA Scoretoolkit (http://publications.idiap.ch/downolads/reports/2012/Anjos_Idiap-Com-02-2012.pdf).

	GMM	GSL	GSL-NAP	GSL-FA	FA	IV-PLDA
Baseline	8.63	8.13	6.31	5.72	5.61	2.98
Replay: office environment	60.32	92.98	29.92	28.54	30.12	30.30
Replay: corridor environment	55.91	88.20	23.59	21.62	24.97	24.53
Replay: anechoic environment	64.40	96.67	49.44	49.31	49.67	49.46
Voice conversion	33.69	36.92	27.58	23.97	23.96	19.30
Speech synthesis	27.29	15.04	13.78	11.91	16.22	10.82

TABLE II: Equal error rates (EERs) for six different ASV systems and for zero-effort impostors (baseline) and three different replay attack configurations (three different acoustic environments). Results are averaged across the three different loudspeaker configurations.

between these two profiles thus serves as an indication of system vulnerability to replay spoofing; in this case the degradation in performance is significant.

This trend is observed across the full set of six ASV systems and nine different replay attack configurations; these results are summarised in Table II. Results in the second row show the baseline performance for each ASV system and for only zero-effort impostors. As expected, the IV-PLDA system delivers the lowest EER. Rows 2–5 illustrate the degradation in performance when zero-effort impostors are replaced with replay attacks applied in three different acoustic environments. EERs in Table II are averaged across the three loudspeaker configurations; other results not reported here showed greater sensitivity to the acoustic environment than to the loudspeaker characteristics. The performance of all six systems degrades significantly. The EER of the most sensitive GSL system increases from 8% to approximately 90% for all three acoustic environments. This degradation is currently attributed to the effect of T-norm, however this hypothesis requires further investigation. Even the EER of the most resistant GSL-FA system increases to between 22% and 50%. Finally, the EER of the state-of-the-art IV-PLDA system increases to between 25% and 50%.

B. Comparison to voice conversion and speech synthesis

DET profiles showing the comparative vulnerabilities to replay, voice conversion and speech synthesis attacks are illustrated in Fig. 6 for the IV-PLDA system. While the comparison of such profiles is

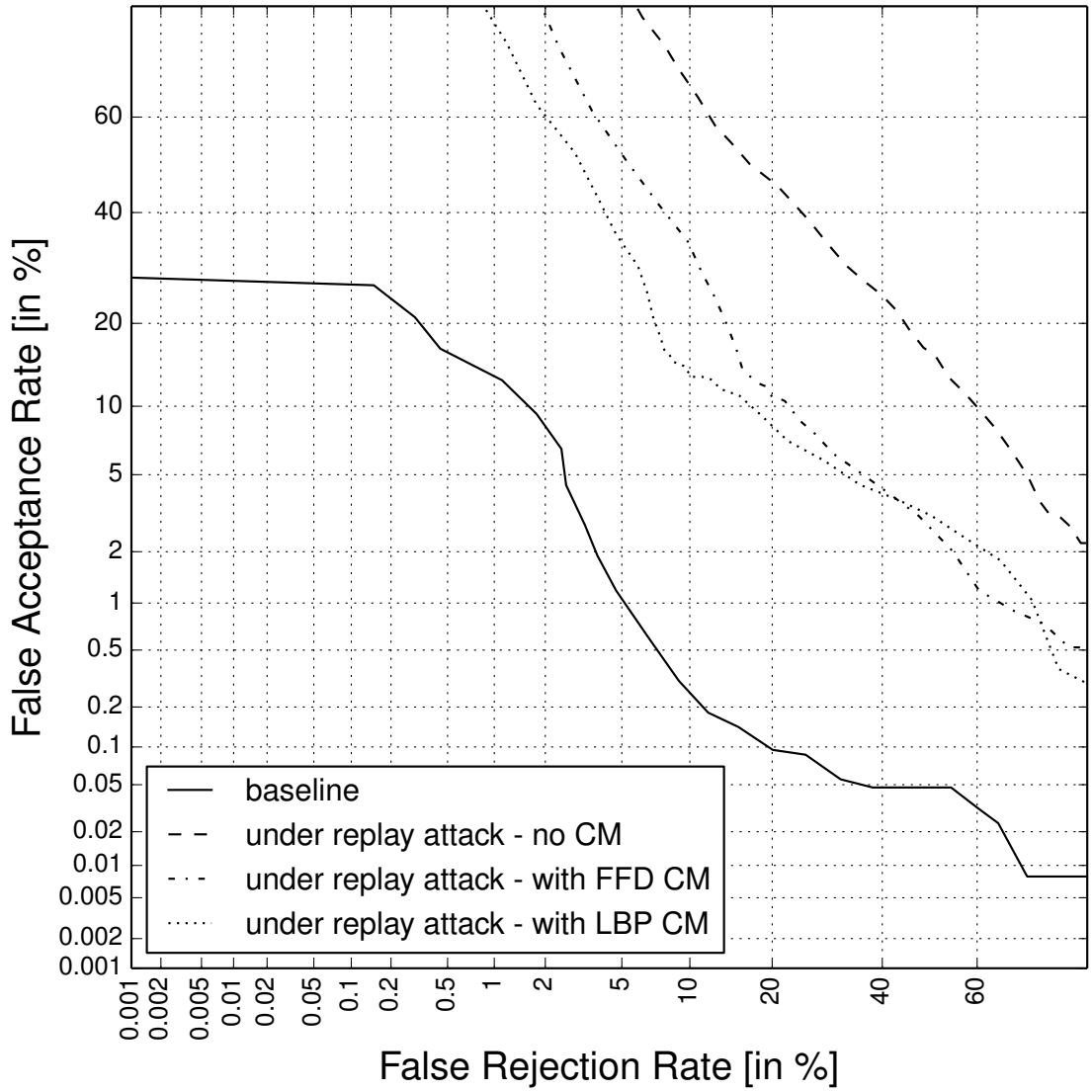

Fig. 5: DET plots for the IV-PLDA system: (i) the baseline, (ii) under replay spoofing with a high-quality speaker in an office, (iii) under replay attack, but with FFD and (iv) LBP countermeasures.

not strictly meaningful[6] it is clear that it is a mistake to discount the threat of replay attacks. This trend is consistent across the full range of ASV systems. Again, while the comparisons are not strictly

[6]Different spoofing algorithms will yield different degradations in ASV performance. For example, other authors [53], show greater vulnerabilities for speech synthesis.

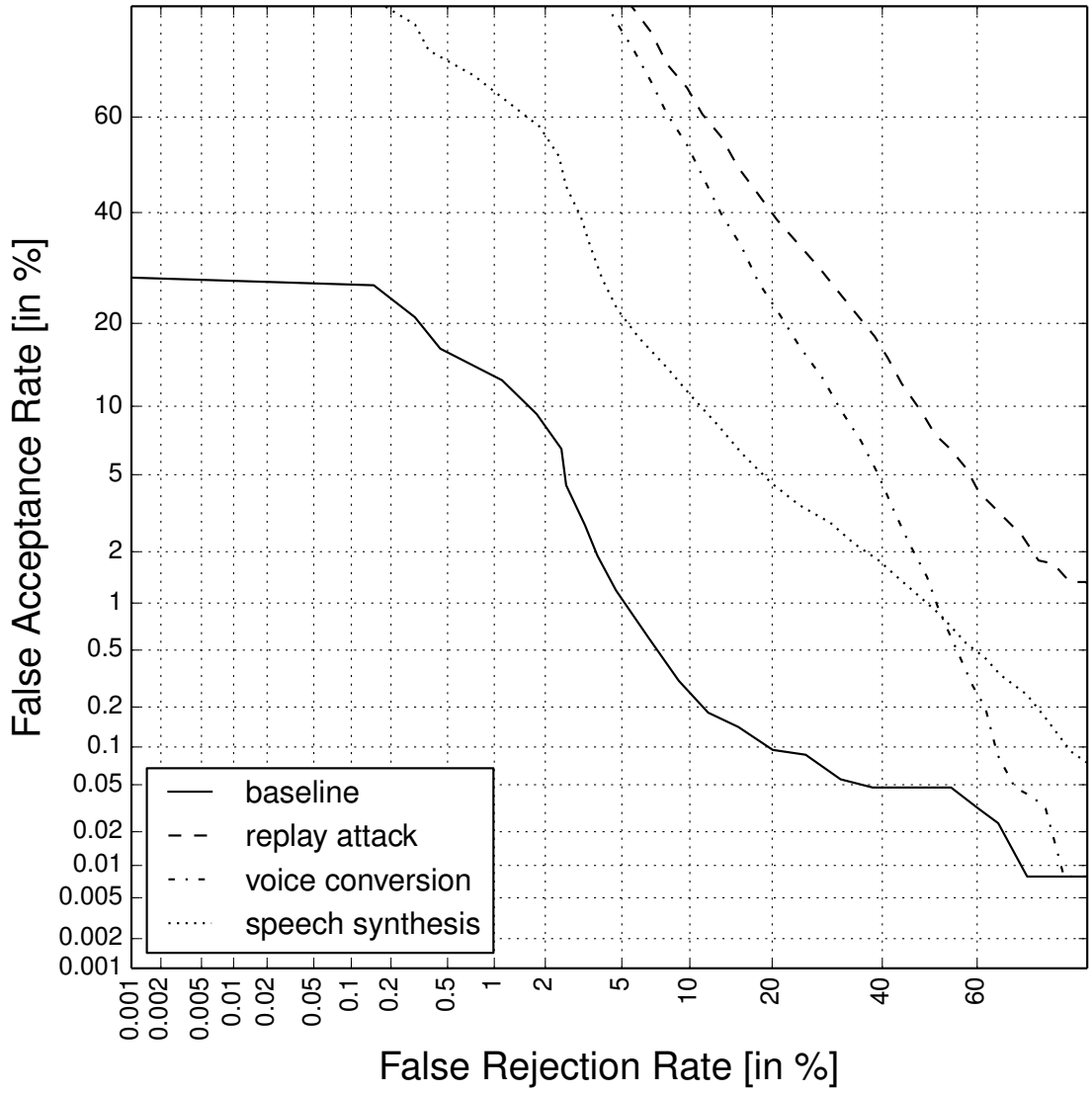

Fig. 6: DET plots for iVector-PLDA system and various attacks.

meaningful, being sufficient only to gauge the relative threat, results in rows 6 and 7 of Tab. II show the broad vulnerability of all six ASV systems to replay attacks.

C. Replay countermeasure

Fig. 7 illustrates (independently of ASV) the performance of the far-field (FFD) and local binary pattern (LBP) countermeasures in detecting replay attacks. Results are illustrated for attacks emulated using a

Environment	EER (%)			SFAR (%)		
	no CM	with FFD	with LBP	no CM	with FFD	with LBP
Office	30.30	13.62	9.56	88.70	63.93	46.29
Corridor	24.53	11.34	7.00	80.91	50.25	30.52
None	49.46	42.14	46.77	97.00	95.17	95.76

TABLE III: EER and SFAR values for various environment of replay attacks, with and without the FFD or LBP countermeasures applied, for IV-PLDA. The SFAR was measured for FRR equal to the baseline EER (2.98%).

high-quality loudspeaker and averaged across both office and corridor environments. Results show that the LBP countermeasure yields better results than the FFD countermeasure (EERs of 2.87% and 16.53% respectively). Since no far-field effects were included in the replay emulation, the comparison is in any case biased in favour of the LBP countermeasure. The step-like nature of the DET curve for the LBP countermeasure is typical in the case of classifiers based on decision trees, such as the AdaBoost M1 classifier used in this work.

The middle two profiles in Fig. 5 show the impact of each countermeasure when integrated with the IV-PLDA ASV system. While both countermeasures show reductions in the EER compared to the upper-most profile, performance still remains far from the baseline.

Detailed results for FFD and LBP countermeasures and three replay environments are presented in Table III. Results are again averaged for the three different loudspeakers and are presented in terms of both EER and SFAR. Unsurprisingly, results for the anechoic environment show that neither countermeasure performs well. Consistent with results in Fig. 5, the LBP countermeasure again outperforms the FFD countermeasure. While replay in office and corridor environments leads to significant degradations in ASV performance, EERs with active countermeasures are less than 10%. Even so, they remain high. Even if the reduction in equivalent SFAR results is significant, and even with active countermeasures, the SFAR remains no lower than 30%. These results suggest that replay spoofing is far from a solved problem.

D. Interpretation of results

The results presented above corroborate the findings of previous work performed with considerably smaller databases; unsurprisingly, replay attacks represent a tangible threat to the reliability of ASV

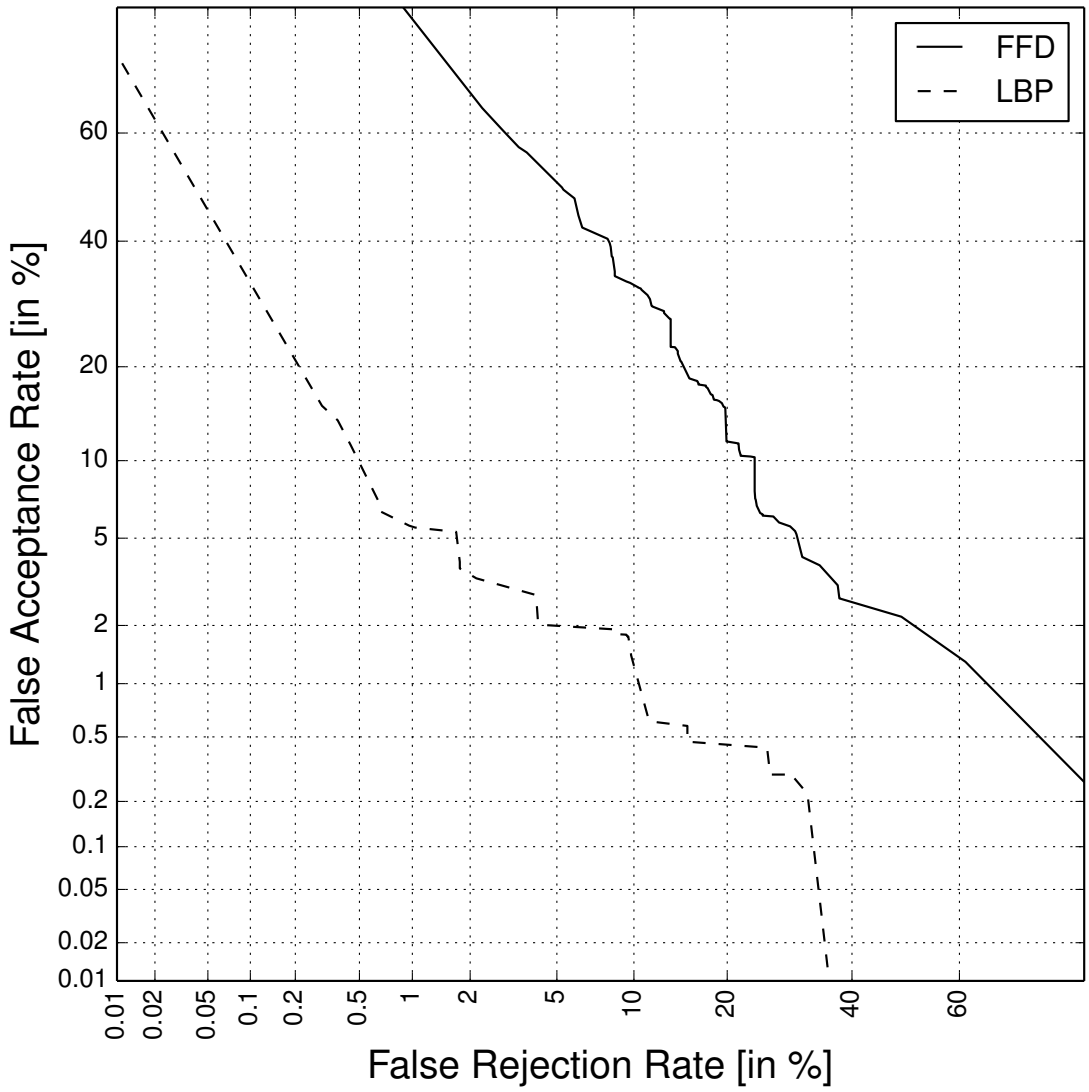

Fig. 7: DET plots for replay spoofing detection using the FFD or LBP countermeasures (assessment independent from ASV), for a high-quality speaker.

system. While reporting the first work with NIST databases and replay attacks emulated in similar fashion to past work with speech synthesis and voice conversion, some aspects of this work must be highlighted in order that results are sensibly and fairly interpreted.

First, the results show that replay attacks represent a threat which is comparable to those of speech synthesis and voice conversion; they are not intended, nor sufficient to show that replay attacks cause a greater degradation in ASV performance. The relative degradations are naturally a function of the authors' effort and expertise in each approach. Other, perhaps more sophisticated speech synthesis and voice conversion algorithms may well cause greater degradations in ASV performance than those reported here. Furthermore, in some sense it is in any case impossible to compare threats on a fully level playing field. Significant application factors, for instance, have not been taken into account in this work. The work should therefore be interpreted only as evidence that the threat of replay to ASV robustness merits greater attention in the future.

These arguments lead naturally to the comparisons regarding the effort and expertise involved in the implementation of each spoofing attack. Whereas voice conversion and speech synthesis are relatively high-effort, high-technology attacks, replay spoofing is implemented easily, without any specific expertise nor sophisticated equipment. Accordingly, the threat posed by replay attacks should be considered greater in a practical perspective, even in the case that future work shows degradations in ASV performance cased by replay are lower than those caused by speech synthesis and voice conversion; they are in any case a significant threat and the most likely form of spoofing 'in the wild'.

VI. Future work

Some aspects of the results require further work to investigate and explain. Of particular note is the vulnerability of the GSL system whose EER increases to 90% when subjected to replay attacks; all other ASV systems have EERs in the order of 25% to 50%. Our initial investigations have shown that score normalisation may be the cause. Other experiments without score normalisation showed EERs of between 30% and 50%. Accordingly, further work is required to study the impact of score normalisation on ASV vulnerabilities to spoofing.

There are also some as-yet-unanswered questions regarding the impact of channel compensation. The main idea behind the detection of replay attacks essentially involves the detection of unexpected channel effects. Thus, while channel compensation has unquestionable utility in ensuring the usability of ASV across different devices, it may also work to the advantage of the fraudster; it can make the very channel characteristics captured by replay countermeasures more difficult to detect. Further work should investigate

this link.

Finally, further work should assess the threat with a more application-driven methodology. That chosen here was motivated by the need to compare the threat of replay to that of voice conversion and speech synthesis, simply as a means of gauging the threat and for prioritising future work. While the NIST SRE datasets were essential in order to support comparisons to other work, and are almost a requirement as regards publication, they were designed more for surveillance and security scenarios rather than authentication applications more relevant to spoofing. Since replay spoofing is undeniably application specific, future work should thus consider more the application than the dataset. It is stressed, however, that this criticism can be levelled equally to all of the past work, including that in speech synthesis and voice conversion spoofing.

VII. Conclusions

This paper assesses the threat to automatic speaker verification of replay spoofing attacks. The threat is shown to be of at least similar significance to that of speech synthesis and voice conversion. The latter have attracted by far the greatest attention in the literature to date whereas replay spoofing is under-researched. This paper argues that replay spoofing merits far greater attention; in contrast to speech synthesis and voice conversion attacks, replay attacks require no specialist expertise nor sophisticated equipment and are therefore the most likely attack in practice.

While there is potential to detect replay spoofing attacks, the approaches assessed in this paper rely upon the presence of measurable channel effects. Approaches to channel compensation, universally popular in today's state-of-the-art speaker verification systems, can reduce channel effects thereby increasing the difficulty in detecting replay attacks. In addition, replay attacks recorded with high-quality sound systems will be almost impossible to detect. Alternative countermeasures will then be needed. Challenge-response mechanisms are a suitable candidate but lack scientific and objective validation.

Lastly, the practical use of these and any other replay countermeasure are fundamentally dependent on the application. It is extremely difficult to assess replay spoofing, hence some rather bold assumptions in this work and the use of replay emulation. While similar assumptions and emulation methodologies typify much of the past work, it is particularly the case with all work related to replay. Further work will be needed to assess replay, and spoofing in general, with an application-driven methodology. Whatever the methodology, however, since any security system is only as strong as its weakest link, countermeasures to thwart replay spoofing merit far greater attention than in the past.

REFERENCES

[1] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech 2013*, Lyon, France, 2013.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communications*, vol. 66, pp. 130–153, 10 2014.

[3] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - a study of technical impostor techniques," in *European Conference on Speech Communication and Technology*, 1999, pp. 1211–1214.

[4] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[5] M. Blomberg, D. Elenius, and E. Zetterholm, "Speaker verification scores and acoustic analysis of a professional impersonator," in *Proc. FONETIK*, 2004.

[6] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "How vulnerable are prosodic features to professional imitators?" in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2008.

[7] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP : Indexation in a Client Memory," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, 2005, pp. 17 – 20.

[8] B. Pellom and J. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, 1999, pp. 837–840.

[9] 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.

[10] P. L. D. Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, march 2010, pp. 1798 –1801.

[11] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Proc. 12th EUSIPCO*, 2012.

[12] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals," in *Proc. 13th Interspeech*, 2012.

[13] Z. Wu, S. Gao, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. APSIPA ASC 2014*, 2014.

[14] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *the Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[15] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[16] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 373–376.

[17] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker adaptive HMM based Text-to-Speech Synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[18] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0," 2007.

[19] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 1985, pp. 5–8.

[20] 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 transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[21] D. Matrouf, J.-F. Bonastre, and J.-P. Costa, "Effect of impostor speech transformation on automatic speaker recognition," *Biometrics on the Internet*, p. 37, 2005.

[22] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[23] ——, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007, pp. 2053–2056.

[24] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. International Conference of the Biometrics Special Interest Group (BIOSIG 2014)*, 2014, (submitted).

[25] D. Petrovska-Delacrétaz and J. Hennebert, "Text-prompted speaker verification experiments with phoneme specific MLPs," in *Proc. ICASSP'98*, 1998.

[26] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 1678–1681.

[27] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, July 2011, pp. 1708–1713.

[28] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, Oct 2011, pp. 1–8.

[29] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[30] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. Interspeech 2013*, Lyon, France, 2013.

[31] The NIST year 2008 speaker recognition evaluation plan. Available: www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf. April 2008.

[32] W. M. Campbell, D. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, may 2006, p. I.

[33] B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *IEEE Transactions on Audio Speech and Language processing*, vol. 15, no. 7, pp. 1960–1968, 2007.

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

[35] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince, "Probabilistic models for inference about identity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 144–157, 2012.

[36] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[37] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, 2010, p. 14.

[38] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42 – 54, Jan 2000.

[39] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, vol. 5, 2008, p. 1.

[40] J.-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "NIST'04 speaker recognition evaluation campaign: new LIA speaker detection platform based on ALIZE toolkit," in *NIST SRE'04*, 2004.

[41] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1155/S1110865704310024

[42] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet, "Overview of the 2000-2001 ELISA consortium research activities," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[43] B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel *et al.*, "Some results from the biosecure talking face evaluation campaign," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4137–4140.

[44] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: an evaluation methodology and lessons learned," *IEEE Signal Processing Magazine, September 2015*, 09 2015.

[45] P. Johnson, B. Tan, and S. Schuckers, "Multimodal fusion vulnerability to non-zero effort (spoof) imposters," in *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*, Dec 2010, pp. 1–5.

[46] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. L. D. Leon, *Speaker recognition anti-spoofing*, S. Marcel, S. Li, and M. Nixon, Eds. Springer, 2014.

[47] M. Wester, "The EMIME bilingual database," The University of Edinburgh, Tech. Rep., 2010.

[48] A. Brown, "Aaron Brown Sound web page," Available: http://www.aaronbrownsound.com/, Apr 2014.

[49] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of the 16th International Conference on Digital Signal Processing*, ser. DSP'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 550–554.

[50] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[51] Y. Freund and R. E. Schapire, "A short introduction to boosting," in *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 1401–1406.

[52] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *in Proc. Interspeech 2015*, Dresden, Germany, 2015.

[53] 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*, vol. 20, no. 8, pp. 2280–2290, 2012.