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 spoofing or presentation attacks in the context of automatic speaker verification (ASV). The implementation of replay attacks requires no specific expertise nor any sophisticated equipment. Replay attacks may thus present a greater risk to ASV than voice conversion or speech synthesis which. As relatively high-technology attacks, the latter are probably beyond the means of the average fraudster. This paper compares the efficacy of each threat under controlled conditions and using seven different ASV systems including a state-of-the-art iVector system with probabilistic linear discriminant analysis. Even if comparatively higher-effort spoofing attacks such as voice conversion and speech synthesis have received greater attention in the recent past, experiments show that low-effort replay attacks provoke greater equal error rates (EERs). Perhaps surprisingly, score normalisation is shown to increase the vulnerability of ASV to replay attacks. 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.**

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 now 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].

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 equipment. Since they are the most easily implemented and while ignoring potential differences in efficacy, it is reasonable to assume that replay attacks will be the most prolific in practice. Nonetheless, the threat of replay attacks has not been quantified using large, standard datasets and hence never compared to that of the comparatively higher-effort attacks which have received considerably greater attention in the literature [2], [18]. With replay attacks being considerably the easiest to implement and with discreet, high quality audio equipment now available to the masses, replay attacks also merit attention.

Only few studies have addressed replay. The work in [3] assessed the vulnerabilities of an HMM-based, text-dependent ASV system with 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 which were collected with close-talk or far-field microphones and then replayed over an analogue or digital telephony channel. The work was conducted with a similarly small corpus with data collected from five speakers and demonstrated the vulnerability of a joint factor analysis (JFA) ASV system; the FAR at the EER threshold increased from 1% to almost 70%. The authors in [19] investigated a text-dependent ASV system exposed to speech replayed using a laptop computer. This first work using the large, standard and publicly available RSR2015 corpus showed that the EER **??what system?? GMM?? iVector??** increased from approximately 4% to more than 20%. **Can you say something further about the thoroughness of the assessments in terms of the number of replay conditions? Do they all, for example, use only a single replay environment? It's about trying to identify the new contribution of this work.**

Missing from the literature, however, is a reliable comparative assessment of replay attacks to voice conversion and speech synthesis using large, standard databases and using a suitably broad range of replay scenarios. Such a study is needed in order to help prioritise future work on developing countermeasures for the greatest threats facing ASV reliability. This paper accordingly aims to assess ASV vulnerabilities to replay attacks using the same ASV systems and base corpora used in previous assessments involving voice conversion and speech synthesis spoofing attacks. 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 2 describes speech synthesis and voice conversion spoofing attacks with a comparison to replay attacks. Section 3 presents previous and ongoing work to develop countermeasures against replay attacks, including our own work using the local binary pattern analysis of speech spectrograms. A common experimental

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.

framework for the assessment of both vulnerabilities and countermeasures is presented in Section 4. Results are presented in Section 5 and our conclusions and ideas for future works are presented in Section 6.

II. Spoofing speaker verification systems

This section summarises our own approaches to assess the vulnerabilities of ASV systems to different forms of spoofing attack and the . In general, all these methods generate a spoofed speech signal $s(t)$ given the speech signal of target speaker $x(t)$. Whereas the input to a speech synthesis system is a text string, that to voice conversion and replay attacks originates from an independent input speech signal $y(t)$.

A. Speech synthesis

There is a large variety of speech synthesis algorithms, such as formant, diphone, unit-selection and statistical parametric 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 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.

Our approach to statistical parametric speech synthesis uses hidden Markov models following the approach described in [12]. Our 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 acceleration coefficients. Acoustic spectral characteristics and duration probabilities are modelled using multispace distribution hidden semi-Markov models (MSD-HSMM) [13]. 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) [14]. 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.

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

B. Voice conversion

An equation for $s(t)$ is missing - suggest simply applying IFFT to equation 2 with simple explanation regarding OLA

Voice conversion has been used to explore ASV spoofing since the late 90s [8], [7]. One of the most successful approaches is so-called Gaussian-dependent filtering approach in [15]. 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 a spoofer $y(t)$ in the spectral domain 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. As such, $y(t)$ is mapped or converted towards the target in a spectral-envelope sense, which is sufficient to overcome most ASV systems.

$H_x(f)$ 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:

$$x_{fil} = \sum_{i=1}^{M} p(g^i_{asr}|y_{asr}) \mu^i_{fil} \tag{3}$$

where $p(g^i_{asr}|y_{asr})$ is the a posteriori probability of Gaussian component g^i_{asr} given the frame y_{asr} and μ^i_{fil} is the mean of component g^i_{fil} which is tied to g^i_{asr}. $H_x(f)$ 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 [15], [17], [16].

C. Replay

Replay attacks are an example of low-effort spoofing; they require simply the replaying of a previously captured speech signal to the ASV microphone. 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 the noise and other intersession (channel) variability compensation techniques. All of these factors point towards the tangible threat posed by replay attacks.

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

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.

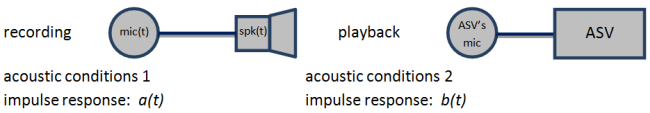

Fig. 1: A schematic diagram of the assumed replay attack configuration. **¡¡¡ Increase font sizes in the figure ¿¿¿**

$$s(t) = x(t) * h(t), \tag{4}$$

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:

$$h(t) = mic(t) * a(t) * spk(t) * b(t) \tag{5}$$

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.

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 is illustrated in Table I, ordered by the level of effort or expertise needed to implement each attack successfully [2].

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 nay 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 effort involved in their implementation; the higher the effort, the greater the impact on ASV performance. However, the proof of concept presented in [20] suggests the contrary, showing that replay attacks pose a high level of risk, being effective in overcoming an ASV system while being the easiest of all concerted-effort spoofing attacks to implement. It is the objective of this paper to investigate these contradictory findings and to compare objectively and quantitatively the comparative 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. Given that only little work has investigated ASV vulnerabilities to such attacks, it is hardly surprising that work to develop anti-spoofing 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 through comparative experiments.

One obvious approach to replay detection involves challenge-response systems which require the speaker to utter a prompted phrase [21]. 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 negative 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 [22]. New access attempts which are deemed too close to previous attempts are rejected. A somewhat similar technique is proposed in [19], 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 were reported in [23] for which the EER for a baseline GMM-UBM system 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.

A. 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 telephone 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 will exhibit far-field channel effects, effects which can be measured and consequently used to detect replayed speech.

This idea was first investigated in [24]. The work compared 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;
- total modulation index, and
- nine sub-band modulation indices – see [24] for precise sub-band bandwidths.

The spectral ratio reflect reflects the level of spectrum flattening or noise and reverberation introduced by far-field recording. The low frequency ratio reflects the 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 showed that far-field recordings could be detected with 90% accuracy. **what classifier?**

B. Local binary patterns

Another method, proposed in [25], uses local binary patterns (LBP) technique. It is based on the hypothesis that modifications made through spoofing disturb the natural 'texture' of genuine speech. This technique was adopted from a standard texture analysis approach, known in image processing [26], to a 2-dimensional 'image' of a speech utterance, where the image is a mel-scaled cepstrogram appended with dynamic features.

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 [26]. This procedure is illustrated in Fig. 2. 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 are determined for each pixel in the 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 parametrisation. The LBP-based countermeasure is based on concatenated histograms formed from the pixel values across each row in the textrogram. The histograms are individually normalised and their resulting bin values are stacked vertically to obtain a new vector in the same manner as GMM mean-vectors are stacked to form supervectors.

This method turned out to be highly effective for artificial signals, yielding 0% EER for the five tested ASVs, very effective for speech synthesis attacks (EER values below 1%) and quite effective for voice conversion – EERs less than 7%. Considering the fact that the LBP-based countermeasure hardly relies on prior knowledge on the attack, in this work we decided to try to use it also as a countermeasure against replay attacks.

IV. Experimental setup

Moved from Section 3.

A. Aim of this work

This paper aims at analysing threat of replay attacks when using large speaker databases and most effective speaker verification systems, and compare it with the threat of other spoofing algorithms: voice conversion and speech synthesis. The impact of acoustic environment and playback devices on performance of speaker verification systems will be also investigated.

Due to lack of real replay recordings (e.g., similar to the MOBIO corpus collected in real environment [27]) we had to use artificial setup of replay environment, however, using impulse responses calculated using real playback hardware and real acoustic environments.

This work will also verify the effectiveness of replay detection, using two previously described replay countermeasures:

- the far-field detector (from here on in referred to as FFD) described in [24], which had been shown to be effective in detecting far-field recordings, and
- the local binary patterns-based detector (from here on in referred to as LBP), described in [25], which yielded satisfactory results for a variety of attacks (artificial signals, speech synthesis and voice conversion).

Therefore, we are going to identify a relative threat of replay using the state-of-the-art ASV systems and two replay countermeasures.

In the following we describe the ASV systems used in this study, the datasets, protocols and metrics, and then the implementation of replay emulation and implementation of the countermeasures.

B. ASV systems

We assessed the impact of each spoofing attacks on seven popular ASV systems: (i) a standard GMM-UBM system with 1024 Gaussian components, as the one used e.g., in [28] (ii) a GMM supervector linear kernel (GSL) system, (iii) a GSL system with nuissance atribute projection (NAP) used for channel compensation [29], (iv) a GSL with factor analysis (FA) [30], (v) a GMM-UBM system with factor analysis, (vi) an iVector system [31], and (vii) an iVector system with probabilistic linear discriminant analysis (PLDA) [32] and length normalisation [33]. A comprehensive comparison of various channel compensation techniques used with GSL kernels can be found, e.g., in [34].

From here on in, the pure iVector system is referred to as IV system, whilst the state-of-the-art iVector system with PLDA is referred to as the IV-PLDA system. All seven ASV systems were tested with and without normalisation. The IV and IV-PLDA systems used symmetric score normalisation (S-norm) as described in [35], while the remaining systems utilised standard T-norm normalisation [36].

All ASV systems used a common speech activity detector which fits a 3-component GMM to the log-energy distribution and which adjusts the speech/non-speech threshold according to the GMM parameters [37]. Such an approach has been used successfully in many independent studies [38], [39] and performed well in comparison to alternatives [40].

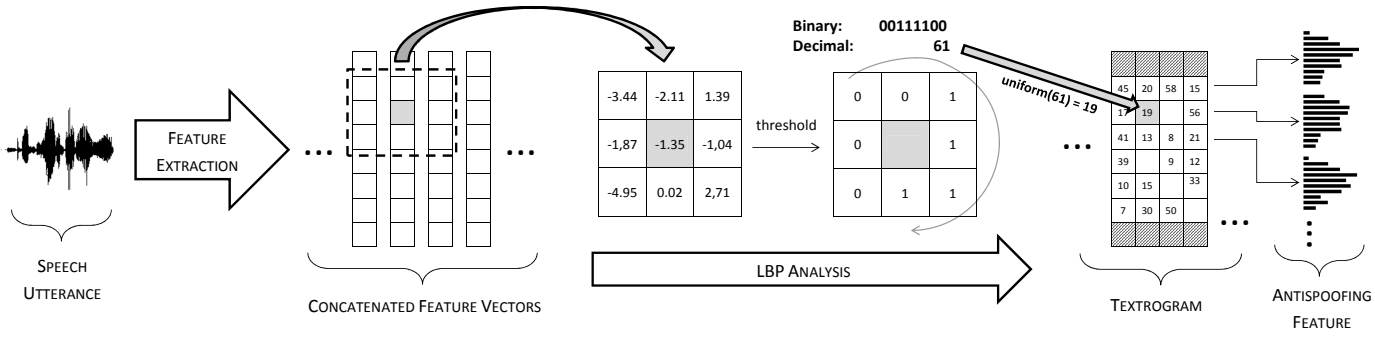

Fig. 2: Schematic diagram of LBP-based feature extraction.

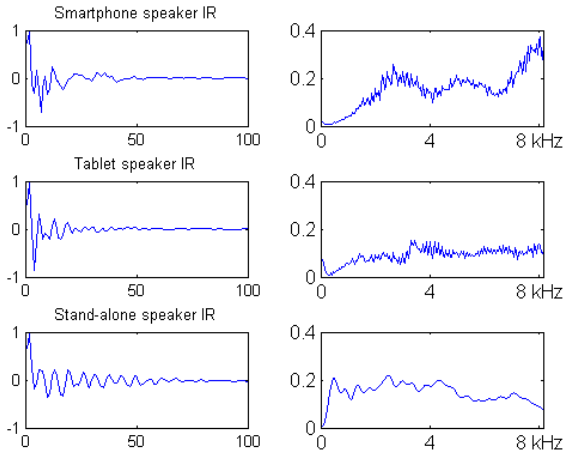

Fig. 3: Impulse responses (left) and corresponding frequency transmittance (right) of the three speakers used for playback emulation.

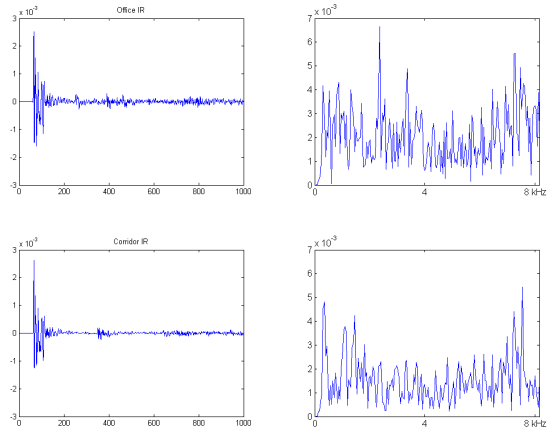

Fig. 4: Impulse responses (left) and corresponding frequency transmittance (right) of the office and the corridor used for playback emulation.

All ASV systems were based on the LIA-SpkDet toolkit [41] and the ALIZE library [42] and were directly derived from the work in [30]. They furthermore used a common UBM with 1024 Gaussian components and a common feature parametrisation: linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy.

C. Datasets, protocols and metrics

All experiments reported below were performed on the male subsets of the 2005 NIST Speaker Recognition Evaluation (NIST'05) and NIST'06 datasets. The former were used for optimising the ASV configurations whereas all results reported later relate to the latter.

In all cases the data used for UBM learning comes from the NIST'04 dataset. Due to the significant amount of data necessary to estimate the total variability matrix T used in the IV-PLDA system, the NIST'06 dataset was additionally used as background data for development whereas the NIST'05 dataset was used as backgroud data for evaluation. In all cases the background datasets were augmented with the NIST'04 and NIST'08 datasets. T is thus learned using approximately 11,000 utterances from 900 speakers, while independence between development and evaluation experiments is always respected.

All experiments related to the 8conv4w-1conv4w condition where one conversation provides 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 was used for enrolment. Experimental results should thus be compared to those produced by other authors 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. In our experiments with replay attacks, all genuine client tests were unchanged, whereas impostor tests were replaced with spoofed (replay) accesses.

Given the consideration of spoofing, and without any specific, standard operating criteria under such a scenario, the equal error rate (EER) is preferred to the minimum detection cost function (minDCF) for ASV assessment. Also reported is the spoofing false acceptance rate (SFAR, [43]) for a false rejection rate (FRR) which is fixed to the EER of the baseline.

D. Voice conversion and speech synthesis setup

The experiments with attacks using voice conversion were conducted with our implementation of the approach originally proposed in [15]. We again considered the worst-case scenario where the attacker/spoofer has full prior knowledge of the ASV system, and so the front-end processing used in voice conversion was exactly the same as that used for ASV. The filtering model and filter $H_x(f)$ used 19 LPCC and LPC coefficients, respectively.

Speech synthesis attacks were implemented using the voice cloning toolkit[2] with a default configuration. We used standard speaker-independent models provided with the toolkit which were trained on the EMIME corpus [44]. The adaptation data for each target speaker comprises three utterances (with transcriptions). Speech signals for spoofing assessment are generated using arbitrary text similar in length to that of true client test utterances.

E. Replay attack setup

Since in this study (as in the majority of other studies on speaker recognition) the NIST databases are used, it implies that the telephony speech will be used. Therefore the $mic(t)$ and $a(t)$ components from the Equation 5 are already determined by the source database, and it can be simplified to:

$$s(t) = x(t) * spk(t) * b(t) \tag{6}$$

To emulate replay attacks at the sensor level we need therefore to reproduce the distortions caused by a replay device ($spk(t)$) with and without the effects introduced by acoustic conditions ($b(t)$). We decided to use the following replay devices:

- a speaker of a popular smartphone brand, from now on in referred to as 'smartphone';
- a speaker of a popular tablet brand, from now on in referred to as 'tablet';
- a stand-alone high-quality speaker, from now on in referred to as 'stand-alone speaker'.

The impulse responses of these speakers are publicly available [45]. Together with frequency responses, they are presented in Fig. 3.

We decided to emulate two likely environments for a spoofing attack: an office room and a corridor. The corresponding impulse responses were taken from the Aachen Impulse Response (AIR) database [46]. We used the impulse response of the office room sized 5.00m x 6.40m x 2.90m, with glass windows, concrete walls, a carpet and typical office furniture, and the impulse response of the corridor sized 18.25m x 2.5m x 2.90m, with concrete walls, five wooden doors and one glass door. To check the impact of replay environment we also run experiments without considering the room acoustics, what would correspond to an anechoic chamber. Impulse and frequency responses of those two rooms are presented in Fig. 4

F. Replay countermeasure setup

1) FFD setup: The FFD countermeasure was set up according to the algorithm proposed in [24].

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

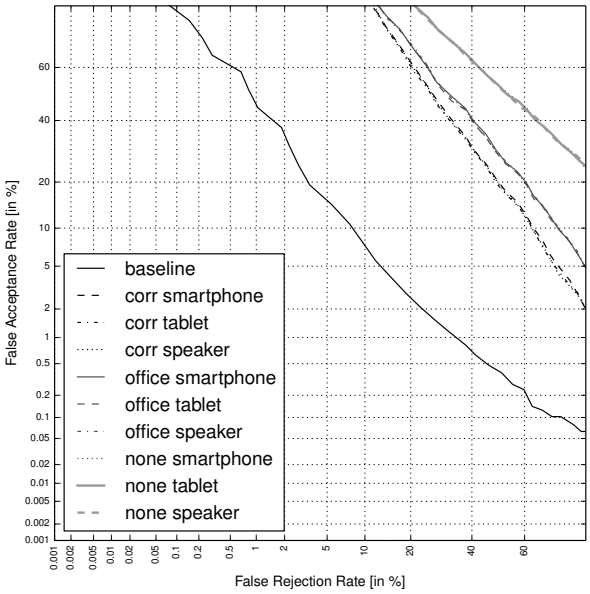

Fig. 5: DET plots for the GMM-UBM system for various replay configurations, compared to the baseline performance.

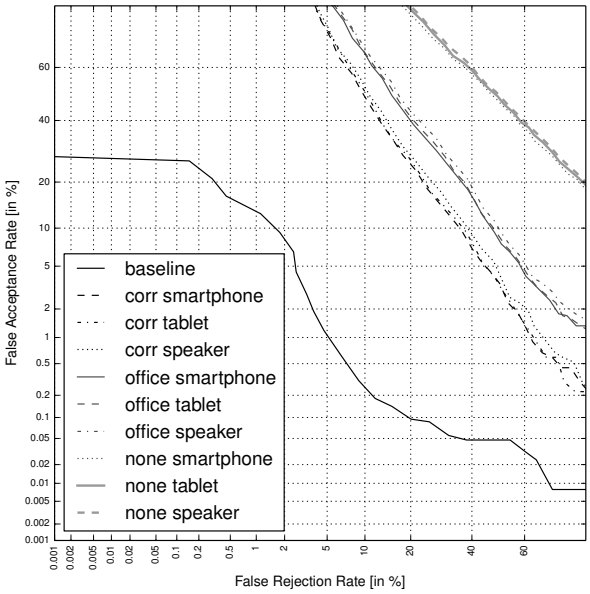

Fig. 6: DET plots for the IV-PLDA system for various replay configurations, compared to the baseline performance.

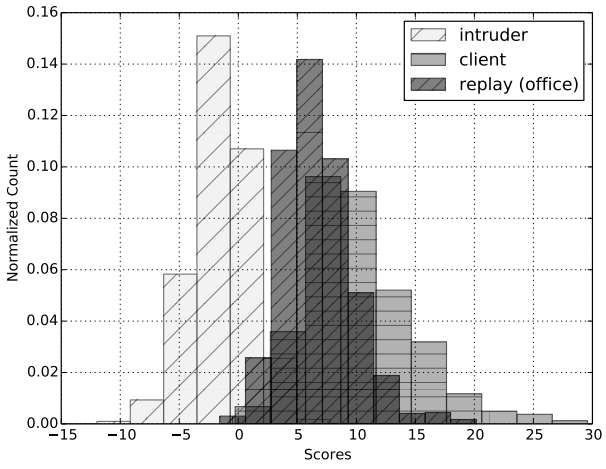

Fig. 7: Score distribution for the IV-PLDA system for replay attacks using a stand-alone speaker and emulation of an office.

The total modulation index was calculated based on the speech signal's envelope. The envelope was approximated by the absolute value of the signal downsampled to 60 Hz. The average modulation index of the signal was calculated for the frames whose index was above a threshold of 0.75.

2) LBP setup: Compared to the original implementation, we reduced the number of possible patterns according to the standard Uniform LBP approach. Uniform LBPs are the subset of 58 patterns which contain at most two bitwise transitions from 0 to 1 or 1 to 0 when the bit pattern is traversed in circularly fashion. As an example, the subset includes patterns 00000001 and 00111100 but not 00110001. As reported by [26], most patterns are naturally uniform and empirical evidence suggests that their use in many image recognition applications leads to better performance then the full set of uniform and non-uniform patterns. We observed similar findings in our previous work [25] and thus decided to ignore pixels corresponding to any of the 198 non-uniform patterns.

We used the implementation made publicly available by The University of Oulu[3]. Normalised features used in the LBP countermeasure were composed of 51 coefficients: 16 LFCCs and energy plus their corresponding delta and delta-delta coefficients. We took into account only those frames determined to contain speech, i.e. those also used for ASV. Histograms of LBPs are created for all but the first and last frames, thereby obtaining a $58 \times 49 = 2842$ length feature vector.

3) Training the replay detectors: Both countermeasure algorithms were trained using a set of 1000 recordings generated using 200 recordings taken from NIST'05 and emulation of various acoustic conditions. In order to make the experiment as close as possible to reality, we decided to use different room impulse responses than the ones used for simulation of replay attack. Therefore we emulated the following environment:

- a lecture room, with concrete walls, glass windows and a parquet;
- a staircase, with concrete walls and steps;
- a meeting room, with concrete walls, glass windows and a carpet.

Similarly to emulation of replay attack, the corresponding impulse responses were taken from the AIR database. To train the far-field recording detector, we also needed to emulate the replay device. We chose a stand-alone speaker impulse response, different from the one used for replay attack emulation, to avoid overfitting to testing data. Original NIST'05 recordings were used to model the licit client access trials.

A binary SVM classifier with polynomial kernel of 3rd degree was used for data classification for the FFD countermeasure, while a classifier based on decision table was used for LBP. Those classifiers returned the best results for those two detectors in terms of the area under the ROC curve. Having been trained, both classifiers were applied to detect replay attacks in both spoofing accesses and licit client trials.

V. Results

Fig. 5 and Fig. 6 present the detection error trade-off (DET) plots[4] for the basic GMM-UBM and the state-of-the-art IV-PLDA systems, exposed to various replay attacks. They show that all replay attacks caused a significant degradation of the ASV performance. However, there are major differences in ASV performance depending on acoustic environment. If the acoustic environment is omitted, the ASV system is almost ideally spoofed – the DET lines are close to straight, with the EER values close to 50%. In realistic cases, when the room acoustics is taken into account, the spoofing is slightly less severe. It can be easily observed that the ASV performance under replay attack in a corridor is better than in an office, probably due to higher level of reverberation for the corridor. Fig. 7 shows a sample score distribution, here for the IV-PLDA system with a replay attack realised in an office (or with speech acquired in an office). A significant overlap of replay attacks with true client accesses can be observed.

In contrast, despite major differences in speakers' impulse responses (see Fig. 3), the differences between the DET plots corresponding to different speakers are only minor. It suggests a relatively low impact of a replay device on the effectiveness of replay attack. Since all seven ASV systems tested showed similar behaviour, therefore, for the sake of clearness, the next results will present the average of the three speakers used in experiments.

Table II shows the detailed EER results for replay attacks in various acoustic conditions against the seven various ASV systems, with and without score normalisation. All systems are shown to be severely sensitive to replay attacks. Even for the highly-reverberant corridor and the most resistant (in terms of EER) GSL kernel system with factor analysis and with T-norm, the EER rose to more than 22% compared to the baseline 5.7%. What was already visible in the DET plots

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

[4]Produced with the TABULA RASA Scoretoolkit (http://publications.idiap.ch/downolads/reports/2012/Anjos_Idiap-Com-02-2012.pdf)

Score norm	Replay env.	GMM	SGL	SGL-NAP	SGL-FA	FA	IV	IV-PLDA
No norm	(Baseline)	9.08	7.89	6.35	6.08	5.60	6.67	3.20
	Office	40.26	34.43	33.52	30.72	33.85	27.83	29.11
	Corridor	35.71	28.24	28.53	25.75	29.92	23.02	22.78
	None	51.59	49.64	49.49	49.73	49.37	49.38	49.37
With norm	(Baseline)	8.63	8.13	6.31	5.72	5.61	6.72	2.98
	Office	60.32	92.98	29.92	28.54	30.12	28.89	30.30
	Corridor	55.91	88.20	23.59	21.62	24.97	23.31	24.53
	None	64.40	96.67	49.44	49.31	49.67	49.06	49.46

TABLE II: EER values for different ASV systems for various acoustic environment of replay attacks, with and without score normalisation.

(see Fig. 6), the results for the office are much worse – the most resistant IV system (without PLDA and without score normalisation) and GSL-FA systems yielded the EER of ca. 28%, whilst the other systems returned EERs of 30% and more. If acoustic conditions are not emulated, the spoofing is almost perfect and the ASV systems yield EER values of 50% or even more.

It is noteworthy that under replay attack the iVector system with probabilistic linear discriminant analysis (PLDA) often shows worse results than the iVector system alone, even though the PLDA significantly decreases the EER for the baseline system (from 6.7% down to less than 3%, with score normalisation). This can be explained that in normal conditions the PLDA improves the performance of iVector-based ASV system as it compensates the intersession differences caused by channel or speaker variation. However, in the case of replay attack this can be disadvantageous, because it also seems to compensate the differences caused by replay devices and replay environments.

A. Comparison of replay threat vs. other spoofing methods

To compare the reply threat with the threat of voice conversion and speech synthesis, we took the average results for all replay devices, as well as the average for office and corridor as the replay environment. The impact of voice conversion, despite demanding considerably more effort to implement, causes a similar degradation in performance to that of replay attacks. E.g., the SGL system yielded the EER of 37% for voice conversion and on average 31% for replay (see Table III); in contrast, the IV-PLDA system showed to be more resistant to voice conversion than replay (20.5% EER vs. 26%, respectively).

High-effort speech synthesis attacks proved much less effective – the EER for the best IV-PLDA system reached less than 11%, whilst replay attack caused an increase of the EER to 20.5%. This confirms the results obtained during preliminary experiments (for smartphone-office combination only) described in [20]. These observations are also illustrated through the DET plot in Fig. 9 for the IV-PLDA system.

B. Impact of score normalisation

The impact of score normalisation in the case of replay attack is ambiguous. For some of the systems, such as factor analysis, GMM supervector linear kernel with factor analysis and with nuisance attribute projection the score normalisation helped to decrease EER values in the face of spoofing, e.g., for the factor analysis system and the corridor the EER decreased from almost 30% to around 25%.

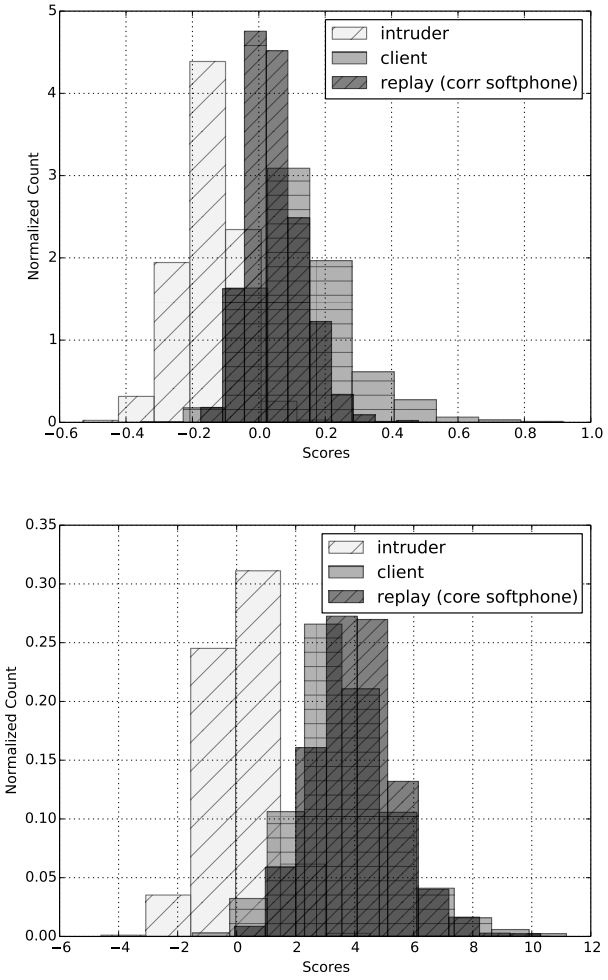

Fig. 8: Score distribution for the GMM-UBM system without (top) and with (bottom) score normalisation.

In contrast, for four other ASV systems (GMM-UBM, GMM supervector linear kernel alone and both iVector systems), the score normalisation in fact helped the spoofer. The increase of EER after applying score normalisation for the GMM-UBM or SGL system is immense, e.g., for replay in an office and the GSL system, the EER increased from 34% to

Attack	GMM	SGL	SGL-NAP	SGL-FA	FA	IV-PLDA
Naïve impostor	9.08	7.89	6.35	6.08	5.60	3.20
Replay	37.99	31.33	31.03	28.24	31.89	25.95
Voice conversion	31.48	36.94	30.44	30.23	23.16	20.45
Speech synthesis	39.90	14.66	13.83	11.98	30.81	10.92

TABLE III: EER values for different ASV systems for various spoofing attacks, without score normalisation.

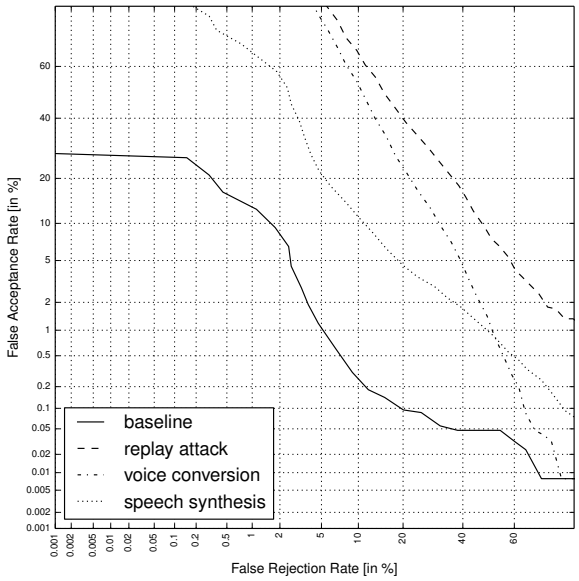

Fig. 9: DET plots for iVector-PLDA system and various attacks.

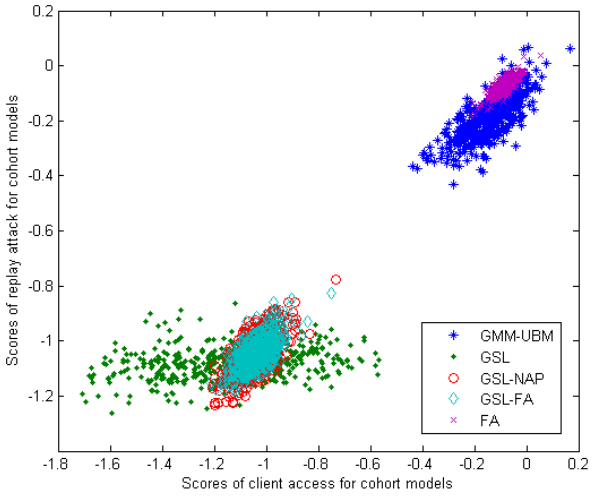

Fig. 10: Score distribution for replay attacks vs. client accesses calculated against cohort speaker models, for various ASVs.

Scores/ASVs	GMM	GSL	GSL-NAP	GSL-FA	FA
Client accesses	0.086	0.252	0.063	0.057	0.032
Replay attacks	0.072	0.060	0.068	0.059	0.027

TABLE IV: Standard deviation of client access scores calculated against cohort speaker models during score normalisation.

93%. This effect is also well illustrated by score distributions presented in Fig. 8, using the example of replay played back from a smartphone in a corridor against a GMM-UBM system. It shows that the scores, having been normalised, got shifted even more to the right than original client accesses.

This phenomenon seems to be a side-effect of T-norm algorithm, which involves dividing the scores by standard deviation of the scores reached for a cohort of reference speaker models. Table IV displays standard deviation values for the client accesses and replay spoofed accesses. It shows that standard deviation for client accesses is by far the largest for the GMM supervector linear kernel system (0.25) and it is also relatively high for the GMM-UBM system (0.09). We believe that this is caused by lack of any compensation mechanism for channel and intersession variability, both in the GMM-UBM and the GSL systems. Higher standard deviation causes shifting the scores of the licit client accesses to the left on the score axis. Contrary, other ASVs cause much lower score dispersion (0.06 or less), what is also visible in Fig. 10.

In contrast, standard deviation of scores for replay attacks is pretty low (0.07 or less, see Table IV) – this makes the normalised scores increase, so this is why they are shifted to the right in Fig. 8, and this is why such systems are so vulnerable to replay attacks. This may pose a significant risk to those ASV systems facing a replay attack.

C. Results of experiments with the replay countermeasure

The detailed results of experiments with FFD and LBP countermeasures for various replay environment, averaged across the replay devices, are presented in Table V. It shows that the countermeasure performance varies depending on acoustic environment. The relative improvement caused by the countermeasures turned out to be the highest for the office – in this case the EER decreased from 30% down to less than 14% for FFD and less than 10% for LBP. Also for the corridor LBP turned out to be more effective than FFD – 7% EER vs. 11%, respectively, also the SFAR result was much lower (30% vs. 46%). When acoustic conditions were not considered, both countermeasures performed poorly, with the EER results slightly better for FFD. This is also visualised by the shape of the DET plots presented in Fig 11.

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 V: 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%).

Environment	EER (%)			SFAR (%)		
	no CM	with FFD	with LBP	no CM	with FFD	with LBP
Smartphone	33.96	11.66	7.52	88.38	54.83	36.79
Tablet	34.48	11.82	7.50	88.42	55.33	36.83
Stand-alone speaker	35.84	13.96	9.83	89.80	61.11	41.60

TABLE VI: EER and SFAR values for various replay devices used for 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%).

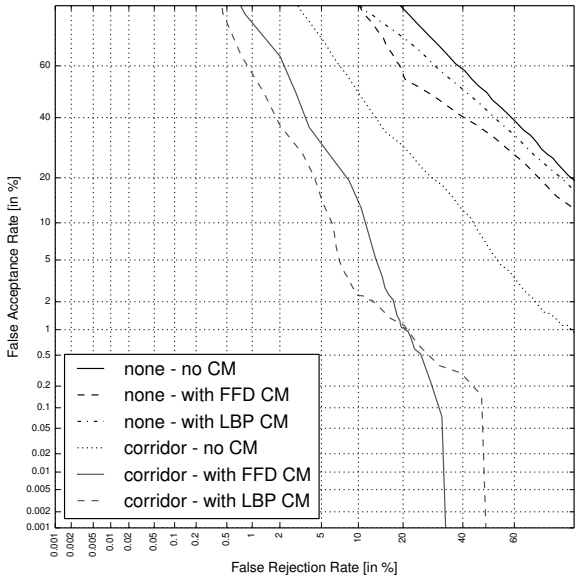

Fig. 11: DET plots for IV-PLDA system for various replay environments, with and without the FFD or LBP countermeasures.

Table VI displays the results of the countermeasure experiments for various replay devices, averaged across different acoustic environments (only office and corridor were taken into account, as they are by far most realistic). Also here the LBP-based countermeasure yields better results than FFD. Both countermeasures helped most for a smartphone and a tablet (the EER values decreased to less than 12% for FFD and to 7.5% for LBP). The results for a stand-alone speaker are only slightly worse (14% and 10%, respectively), most likely due to higher quality of this device (see much better frequency response shown in Fig. 3).

VI. Conclusions

Despite the lack of attention to replay attacks in the literature, results show that low-effort replay attacks pose a significant risk, surpassing that of comparatively high-effort attacks such as voice conversion and speech synthesis. The important contribution of the presented work is the conclusion that the techniques which normally improve the performance of ASV systems, such as score normalisation or PLDA (for iVectors), in fact can work in favour of the spoofer.

We have also showed that the proposed countermeasure based on local binary patterns can be quite effective in detecting replay attacks, outperforming in realistic acoustic conditions the far-field recording detection described in [24]. However, it must be stressed that an attack using high-quality recordings (e.g., acquired in an anechoic booth) can be very difficult to detect.

Given that the implementation of replay attacks demands neither specific expertise nor any sophisticated equipment, the risk to ASV is arguably greater than that of voice conversion and speech synthesis which currently receive the most attention in the literature. Future evaluation should not only consider the threat of any particular attack, but also the ease with which they can be performed. We suggest that a risk-based approach should be adopted.

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, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals," in *Proc. 13th Interspeech*, 2012.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007, pp. 2053–2056.

[17] ——, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[18] 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.

[19] 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.

[20] 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).

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkcnafet, L. Mazaira Fernandez, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Gunther, J. Zganec-Gros, R. Candil, F. Simoes, M. Bengherabi, A. Alvarez Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz, D. Van Leeuwen, J. Gonzalez-Dominguez, M. Neto, E. Boutellaa, P. Gomez Vilda, A. Varona, D. Petrovska-Delacretaz, P. Matejka, J. Gonzalez-Rodriguez, T. Pereira, F. Harizi, L. Rodriguez-Fuentes, L. El Shafey, M. Angeloni, G. Bordel, G. Chollet, and S. Marcel, "The 2013 speaker recognition evaluation in mobile environment," in *Biometrics (ICB), 2013 International Conference on*, June 2013, pp. 1–8.

[28] A. Roy, M. Magimai-Doss, and S. Marcel, "A fast parts-based approach to speaker verification using boosted slice classifiers," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 1, pp. 241–254, Feb 2012.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 4, pp. 802–809, Dec 2010.

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

[36] 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.

[37] 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

[38] 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.

[39] 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.

[40] M. Sahidullah and G. Saha, "Comparison of speech activity detection techniques for speaker recognition," *arXiv e-preprint*, Oct. 2012.

[41] 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.

[42] 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.

[43] 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.

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

[45] A. Brown, "Aaron brown sound web page," http://www.aaronbrownsound.com/, Apr 2014.

[46] 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.

Artur Janicki Biography text here.

Federico Alegre Biography text here.

Nicholas Evans Biography text here.