This Will We the Title of the Paper for the IEEE Transactions on Information Forensics and Security

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—This paper re-examines the threat of spoofing or presentation attacks in the context of automatic speaker verification (ASV). While voice conversion and speech synthesis attacks present a serious threat, and have accordingly received a great deal of attention in the recent literature, they can only be implemented with a high level of technical know-how. In contrast, the implementation of replay attacks require no specific expertise nor any sophisticated equipment and thus they arguably present a greater risk. The comparative threat of each attack is re-examined in this paper against six different ASV systems including a state-of-the-art iVector-PLDA system. Despite the lack of attention in the literature, experiments show that low-effort replay attacks provoke higher levels of false acceptance than comparatively higher-effort spoofing attacks such as voice conversion and speech synthesis. Results therefore show the need to refocus research effort and to develop countermeasures against replay attacks in future work.**

***Index Terms*—IEEEtran, journal, LaTeX, paper, template.**

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 acceptance. 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 impersonation [3], [4], voice conversion [5], [6], speech synthesis [7], [8] and attacks with non-speech, artificial tone-like signals [9].

Common to the bulk of previous work is the consideration of attacks which require either specific skills, e.g. impersonation, or high-level technology, e.g. speech synthesis and voice conversion. With the noteworthy exceptions of [10], [11], relatively little attention has been paid to low-effort spoofing attacks such as replay. Replay attacks can be performed without any specific expertise nor any sophisticated equipment. Since they are the most easily implemented, it is natural to assume that replay attacks will be the most commonly encountered in practice. Nonetheless, the threat of replay attacks has neither been quantified using large, standard datasets nor compared to that of voice conversion or speech synthesis attacks. This paper accordingly aims to re-assess ASV vulnerabilities to replay attacks using the same ASV systems and corpora used in previous assessments involving voice conversion and speech synthesis spoofing attacks. The results of this contribution are in contrast to our original hypothesis that lower effort spoofing attacks are less effective.

M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised December 27, 2012.

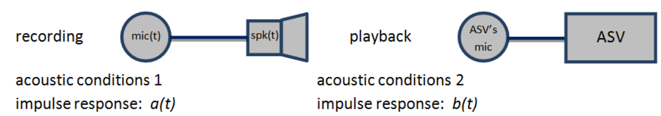

Fig. 1: Schematic diagram of replay.

The paper is organised as follows. Section 2 describes an approach to simulate replay attacks in order that their effect can be compared to those of voice conversion and speech synthesis using the same corpora. A common experimental setup in which the vulnerabilities of six different ASV systems is presented in Section 3. Results are presented in Section 4 and our conclusions and ideas for future work are presented in Section 5.

II. Replay vs. other spoofing methods

Replay is an example of low-effort spoofing attacks; they require simply the replaying of a previously captured speech signal. Replay attacks can be realised with increasing ease, considering the widespread availability of mobile devices with reasonable quality in-built speakers (and microphones). The risk of playback attacks is even higher if recordings of a speaker are publicly available, and if text-independent system is used. Paradoxically, increased effectiveness of channel-compensation techniques and methods for compensating intersession variability can actually work in favour of the replay attack. All these factors show the increasing threat of replay attack on speaker verification systems and justify the importance of research on replay and replay countermeasures.

When modelling a replay attack one should take into account the impact of the following elements:

- acoustic effects introduced by the recording device;
- acoustic conditions in the environment where the voice was acquired;
- acoustic effects of the replay device, and the
- acoustic conditions in the environment where the attack takes place.

If $x(t)$ is the speech signal of the client, the playback (spoofing) signal $y(t)$ can be represented by:

$$y(t) = x(t) * mic(t) * a(t) * spk(t) * b(t) \quad (1)$$

where * denotes convolution, $mic(t)$ and $spk(t)$ are impulse responses of the microphone and the speaker, respectively, and

$a(t)$ and $b(t)$ are impulse responses of recording and replay environments, respectively (see Fig. 1).

A. Research on replay spoofing and replay countermeasures

While a great deal of attention has been paid to medium- and high-effort spoofing algorithms (a thorough review of these can be found, e.g., in [1]), surprisingly, only few studies have been published so far on replay spoofing. The work in [10] assessed the vulnerabilities of an HMM-based text-dependent ASV system with concatenated digits. They showed that replay attacks are highly effective, but their experiments related to only two speakers. In the study of [11] several playback cases were analysed: recording using a close-talk or a far-field microphone and transmission over an analogue or digital channel. Using their own corpus with five speakers the work showed that a joint factor analysis (JFA) ASV system is vulnerable to replay attacks – the FAR at the EER threshold increased from 1% to almost 70%.

Some measures were also proposed to prevent replay attacks. As far as text-dependent ASV systems are concerned, one of them is the use of challenge-response systems, which require the speaker to utter an ad hoc phrase which may not be easily expected. Another method which was developed for text-dependent systems (but may be easily adopted for text-independent systems) is based on comparing a new access trial with stored previous attempts [12]. The experiments showed that this method caused a decrease in EER in most of the cases, however, this method is useless if there were no previous access trials with a given recording. In other work [13] measuring of the channel noise is proposed in order to find a difference between the expected (simpler) channel and the more noisy replay channel, which includes also the recording device. The authors proposed two variants of their countermeasure, and they managed to decrease the EER from 40% to 10% with a baseline GMM-UBM system under replay spoofing.

Villalba and Lleida [14] proposed a countermeasure which was based on detecting far-field recordings, bearing in mind that both telephone-based ASV systems and stand-alone ASV systems (e.g., installed at the entrance to a room) expect close-talk speech. The authors observed that far-field recordings cause changes in speech signal envelope; therefore they extracted 12 parameters describing the envelope and, based on them, trained a binary SVM classifier in order to discriminate far-field speech from close-talk speech. The authors claimed that they reached the far-field recognition rate of more than 90%.

B. Replay vs. other spoofing algorithms

Table I shows a comparison of replay spoofing and naïve (zero-effort) impostors, voice conversion and speech synthesis. The attacks are ordered in terms of the effort involved in each case. Replay attacks require slightly increased effort compared to naïve imposture (need for target voice acquisition and replay hardware). Voice conversion and speech synthesis require specialised algorithms, in addition to appropriate hardware and parameters describing the target voice. They belong to a class of higher-effort spoofing attacks. While voice conversion is still based upon the conversion of an original speech signal, speech synthesis starts with text input. In this sense the attack requires the most effort of all to implement successfully. One may reasonably suppose that the effectiveness of each attack is linked to the effort involved; the higher the effort, the greater the impact on ASV performance. This hypothesis will be verified later in this work.

C. Voice conversion

We used the approach to voice conversion originally presented in [15]. At the frame level, the speech signal of a spoofer denoted by $y(t)$ is filtered in the spectral domain as follows:

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

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 whereas $Y'(f)$ 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 whereas 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:

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

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 [15], [16], [6].

D. Speech synthesis

There is a large variety of speech synthesis algorithms, such as formant, diphone or unit-selection based synthesis. State-of-the-art text-to-speech systems use either unit-selection or the hidden Markov model-based synthesis (HTS). Whilst the former requires large amounts of speech data, the latter does not, and can therefore much more easily generate speech targeted towards a specific client.

Accordingly, in this paper we consider spoofing with HTS synthesis, following the approach described in [17], and using the HMM-based Speech Synthesis System (HTS)[1]. Parametrisation includes STRAIGHT (Speech Transformation

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

Attack	Naïve impostor	Replay	Voice conversion	Speech synthesis
Speech used	impostor's (genuine)	client's	impostor's (converted)	synthetic
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.

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) [18]. 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) [19]. 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.

E. 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 investigated.

This work will also verify the effectiveness of replay detection, using the previously described replay countermeasure [14], which had been shown to be effective in detecting far-field recordings. Therefore, we are going to identify a relative threat of replay using the state-of-the-art ASV systems and a replay countermeasure.

III. Experimental setup

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 countermeasure, based on detection of far-field recordings.

A. 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, (ii) a GMM supervector linear kernel (GSL) system, (iii) a GSL system with nuissance atribute projection (NAP) used for channel compensation [20], (iv) a GSL with factor analysis (FA) [21], (v) a GMM-UBM system with factor analysis, (vi) an iVector system [22], and (vii) an iVector system with probabilistic linear discriminant analysis (PLDA) [23] and length normalisation [24].

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 [25], while the remaining systems utilised standard T-norm normalisation [26].

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 [27]. Such an approach has been used successfully in many independent studies [28], [29] and performed well in comparison to alternatives [30].

All ASV systems were based on the LIA-SpkDet toolkit [31] and the ALIZE library [32] and were directly derived from the work in [21]. 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.

B. 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.

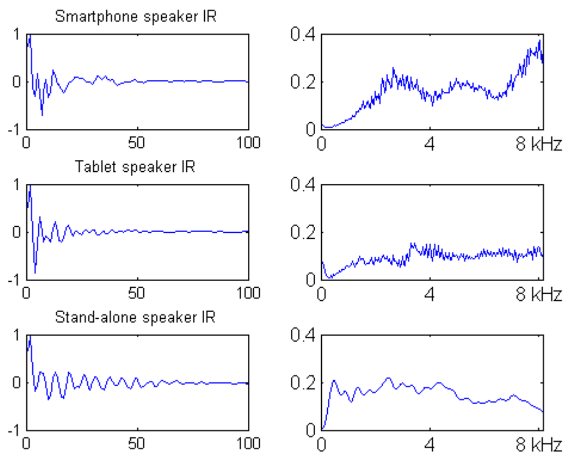

Fig. 2: Impulse responses (left) and corresponding frequency transmittance (right) of the three speakers used for playback emulation.

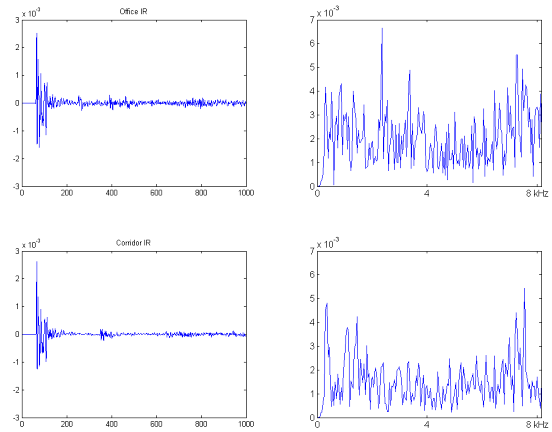

Fig. 3: Impulse responses (left) and corresponding frequency transmittance (right) of the three speakers used for playback emulation.

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 false acceptance rate (FAR) for a false rejection rate (FRR) which is fixed to the EER of the baseline.

C. 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 1 are already determined by the source database, and the Equation 1 can be simplified to:

$$y(t) = x(t) * spk(t) * b(t) \tag{4}$$

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 [33]. Together with frequency responses, they are presented in Fig. 2.

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 [34]. 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.

D. Replay countermeasure setup

The countermeasure was set up according to the algorithm proposed in [14]. Each recording, both from the training and the testing datasets, was additionally described using the following 12 parameters:

- spectral ratio – the ratio between the signal energy from 0 to 2 kHz and from 2 kHz to 4 kHz;
- low frequency ratio – ratio between the signal energy from 100 Hz to 300 Hz and from 300 Hz to 500 Hz;
- total signal modulation index; and
- nine sub-band modulation indices, for sub-bands: 1kHz-3kHz, 1kHz-2kHz, 2kHz-3kHz, 0.5kHz-1kHz, 1kHz-1.5kHz, 1.5kHz-2kHz, 2kHz-2.5kHz, 2.5kHz-3kHz and 3kHz-3.5kHz.

The countermeasure algorithm was 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

Score norm	Replay env.	GMM	SGL	SGL-NAP	SGL-FA	FA	IV	IV-PLDA
	(Baseline)	9.08	7.89	6.35	6.08	5.60	6.67	3.20
No	Office	40.26	34.43	33.52	30.72	33.85	27.83	29.11
norm	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
	(Baseline)	8.63	8.13	6.31	5.72	5.61	6.72	2.98
With	Office	60.32	92.98	29.92	28.54	30.12	28.89	30.30
norm	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.

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. After being trained with the training data in 12-dimensional space defined by the parameters described above, it was applied to detect replay attacks in both spoofing accesses and licit client trials.

E. Setup of other spoofing attacks

The experiments with attacks using voice conversion were conducted with our implementation of the approach originally proposed in [15]. We again consider 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 [35]. 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.

IV. Results

Fig. 4 and Fig. 5 present the detection error trade-off (DET) plots[3] 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.

These results are confirmed by score distribution presented in Fig. 6 – they show that without considering the acoustic environment (the bottom plot) the scores for replay attacks overlap the scores of genuine client accesses, whilst the scores returned by the IV-PLDA system with a replay attack realised in a corridor (or with speech acquired in a corridor) are shifted towards impostor trials, however still with a significant overlap with true client accesses (see the middle plot of Fig. 6).

In contrast, despite major differences in speakers' impulse responses (see Fig. 2), 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 (see Fig. 5), 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.

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; in contrast, the IV-PLDA system showed to be more resistant to voice conversion than replay (20.5% EER vs. 26%, respectively).

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

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

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.

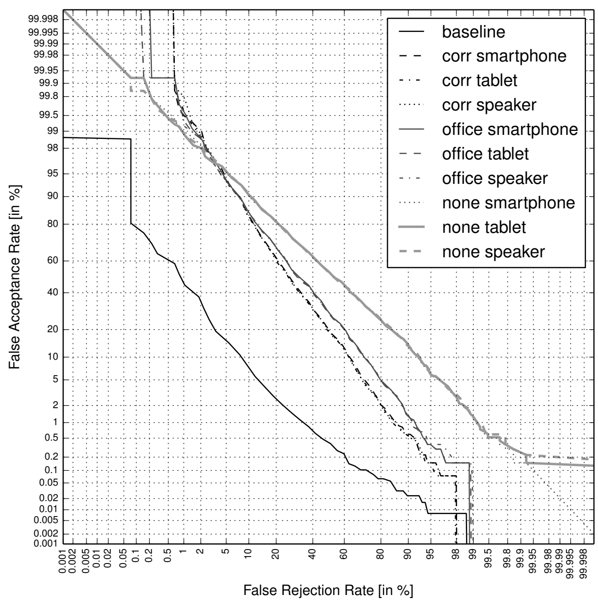

Fig. 4: DET plots for the GMM-UBM system for various replay configurations, compared to the baseline performance.

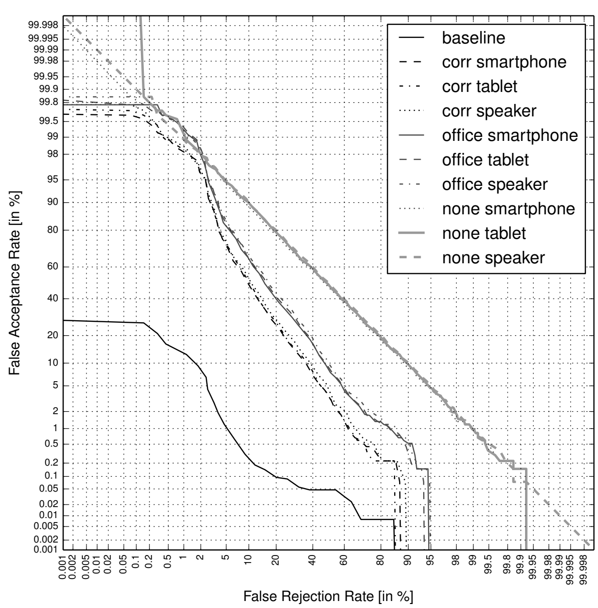

Fig. 5: DET plots for the IV-PLDA system for various replay configurations, compared to the baseline performance.

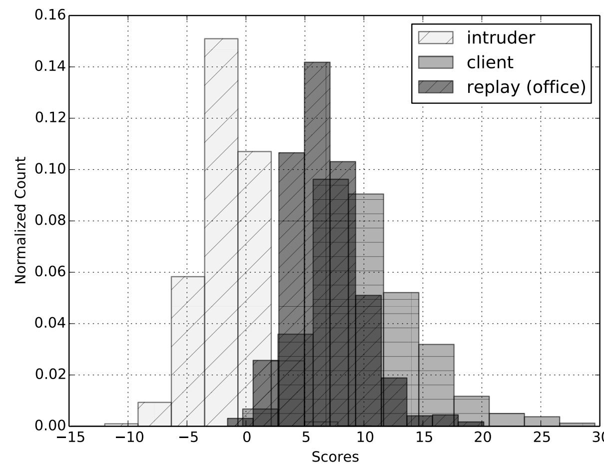

Fig. 6: Score distribution for the IV-PLDA system for replay attacks using a stand-alone speaker and with simulation of an office.

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%. These observations are also illustrated through the DET plot in Fig. 8 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%.

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 93%. This effect is also well illustrated by score distributions presented in Fig. 7, 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

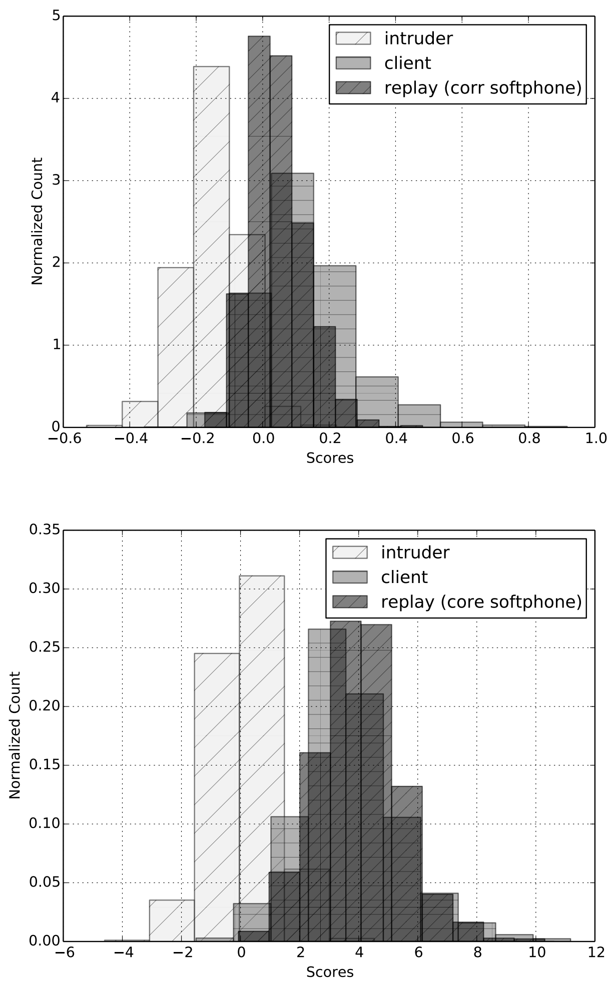

Fig. 7: Score distribution for the GMM-UBM system without (top) and with (bottom) score normalisation.

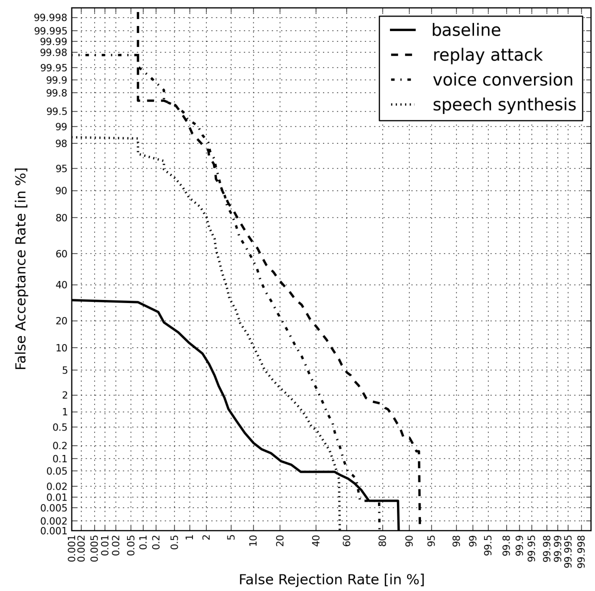

Fig. 8: DET plots for iVector-PLDA system and various attacks.

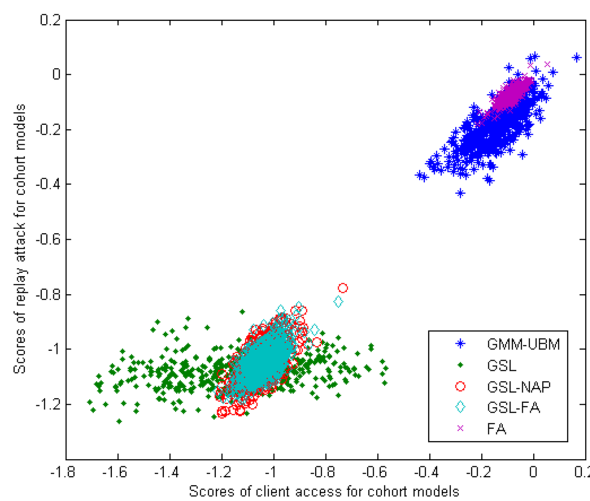

Fig. 9: Score distribution for replay attacks vs. client accesses calculated against cohort speaker models, for various ASVs.

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.

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. 9.

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. 7, 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. Impact of using PLDA with iVector systems

When analysing operation of various ASV systems, we also compared the performance of both iVector systems used in

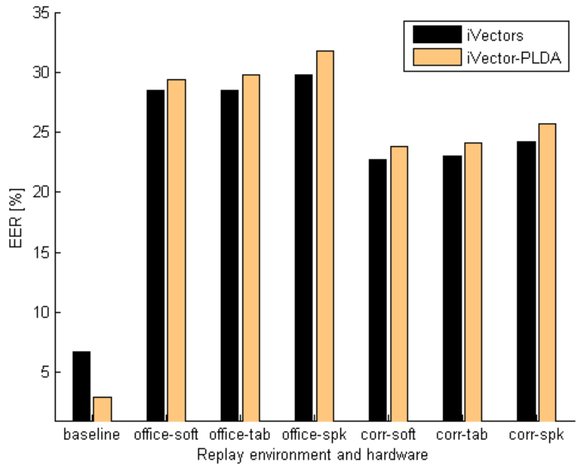

Fig. 10: EER results for iVector-based ASV system with and without PLDA, with score normalisation used.

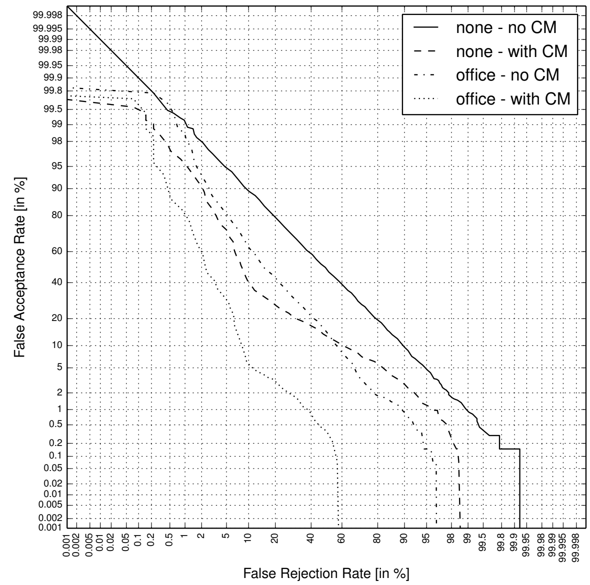

Fig. 11: DET plots for IV-PLDA system for various replay environments, with and without the countermeasure.

experiments - with and without probabilistic linear discriminant analysis (PLDA). The results in Table II, confirmed by Fig. 10, clearly show that even though the PLDA significantly decreases the EER for the baseline system (from 6.7% down to less than 3%, with score normalisation), it mostly increases the EER when under replay attack. E.g., for an office and score normalisation the EER rose from less to 29% to 30.3%.

This behaviour of PLDA can be explained as follows: in normal conditions the PLDA improves significantly 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.

D. Results of experiments with the replay countermeasure

The detailed results of experiments with far-field detection countermeasure 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 countermeasure turned out to be the highest for the most difficult case, i.e., when the acoustic conditions are not taken into account – in this case the EER decreased from almost 50% down to 30%. The relative improvement was the lowest for the case where spoofing was initially least successful, i.e. for the corridor – it decreased from 24% to less than 14%. Anyway, after applying the countermeasure, it was for the corridor where the absolute EER value was the lowest, what we believe is caused by the high reverberation of the corridor.

Table VI, in contrast, displays the results of the countermeasure experiments for various replay devices, averaged across different acoustic environments. Even though surprisingly the reply devices had no major impact on spoofing effectiveness (all three devices made the IV-PLDA system yield the EER of ca. 35%), they do have strong impact of effectiveness of reply countermeasure. The far-field detection countermeasure seems to help best for a smartphone and slightly worse for a tablet (the EER values decreased to less than 12% and 15%, respectively). However, the countermeasure worked poorly for a stand-alone speaker – the EER decreased here on average only by ca. 3%.

These results may seem somewhat surprising, considering the fact it was the IR of a stand-alone speaker which was used to train the far-field recording detector, while neither any smartphone IR nor any tablet IR were used to create the training corpus. Anyway we believe that the poor detection of replay attacks caused by the stand-alone speaker is caused by high quality of audio playback provided by such speakers (see e.g. much better frequency response for a stand0alone speaker shown in Fig. 2), what obviously makes reply detection much more difficult (less than one third of attacks detected if no acoustic conditions considered, at the level of 11.5% of false rejections).

Even though the countermeasure applied detected more than 90% of replay attacks when using a smartphone and a tablet (at the level of 11.5% of false rejections), and the EER and also FAR values decreased significantly, the FAR values at the false rejection rate set to baseline EER (2.98%) remain still high (e.g., 49% for a smartphone and 54% for a tablet). This is also confirmed by the shape of DET plots presented in Fig 11. This, however, can be modified within certain limits by adjusting the threshold of the far-field recording classifier.

V. Conclusions

Despite the lack of attention to replay attacks in the literature and contrary to our hypothesis, 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. Worthy of note is the performance of the

Environment	% attacks detected	EER no CM	EER with CM	FAR no CM	FAR with CM
Office	80.46	30.30	14.46	88.70	52.88
Corridor	73.09	24.53	13.82	80.91	48.02
None	62.90	49.46	30.03	97.00	86.30

TABLE V: Replay attack recognition rate, EER and FAR values for various environment of replay attacks, with and without the countermeasure applied. The FAR was measured for FRR equal to the baseline EER (2.98%), replay recognition rate given for false positive rate of 11.53%.

Environment	% attacks detected	EER no CM	EER with CM	FAR no CM	FAR with CM
Smartphone	96.01	33.96	11.47	88.38	49.31
Tablet	87.82	34.48	14.58	88.42	54.93
Stand-alone speaker	32.61	35.84	32.25	89.80	82.96

TABLE VI: Replay attack recognition rate, EER and FAR values for various replay devices used for attacks, with and without the countermeasure applied. The FAR was measured for FRR equal to the baseline EER (2.98%), replay recognition rate given for false positive rate of 11.53%.

state-of-the-art iVector-PLDA system which, despite showing the best baseline performance, is the most vulnerable to replay attacks, especially for FARs below 10%.

Future work should thus pay greater attention to replay attacks and, in particular, suitable replay attack countermeasures. The assumption that higher-effort attacks pose the greatest threat might be ill-founded. 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] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[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] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normaliza-

tion for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42 – 54, Jan 2000.

[27] 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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

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

John Doe Biography text here.

Jane Doe Biography text here.