Re-assessing the threat of replay spoofing attacks against automatic speaker verification

Federico Alegre
EURECOM
Sophia Antipolis, France
alegre@eurecom.fr

Artur Janicki *
Warsaw University of Technology
Warsaw, Poland
A.Janicki@tele.pw.edu.pl

Nicholas Evans
EURECOM
Sophia Antipolis, France
evans@eurecom.fr

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.

1 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 [EKY13, EKY$^+$14]. A growing body of independent work has now demonstrated the vulnerability of ASV systems to spoofing through impersonation [FWAH08], voice conversion [PAB$^+$05, BMF07], speech synthesis [MHTK99, LAPY10] and attacks with non-speech, artificial tone-like signals [AVE12].

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

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

voice conversion. With the noteworthy exceptions of [LB99, VL10], 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.

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.

2 Replay attacks vs. high-effort spoofing

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.

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:

$$y(t) = x(t) * mic(t) * a(t) * spk(t) * b(t) \tag{1}$$

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. In this study we consider the worst-case scenario, in which the spoofer possesses high quality recordings of the client. The impact of the

recording device and recording environment room can thus be neglected and Equation 1 is simplified to:

$$y(t) = x(t) * spk(t) * b(t) \tag{2}$$

Surprisingly, only few studies have been published so far on replay spoofing. The work in [LB99] 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 [VL10] 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%.

In contrast, 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 [EKY13]. They typically used large corpora (such as the NIST databases). This paper aims to investigate the threat of replay attacks with large databases and to compare the effectiveness of replay spoofing with the most effective medium- and high-effort spoofing algorithms - voice conversion and speech synthesis. These two attacks are described in the following.

2.1 Voice conversion

We used the approach to voice conversion originally presented in [MBC05]. At the frame level, the speech signal of a spoofer denoted by $y(t)$ is filtered in the spectral domain as follows:

$$Y'(f) = \frac{|H_x(f)|}{|H_y(f)|} Y(f) \tag{3}$$

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:

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

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	**(?)**	high	high

Table 1: Comparison of four different attacks in terms of speech used, required effort and effectiveness.

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 [MBC05, BMF06].

2.2 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 [YNZ+09], and using the HMM-based Speech Synthesis System (HTS)[1]. Parametrisation includes 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) [RM85]. 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) [YKN+09]. 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/

2.3 Replay vs. voice conversion and speech synthesis

Table 1 shows a comparison of naïve (zero-effort) impostors to replay, voice conversion and speech synthesis, as well as with naïve imposture. The attacks are ordered in terms of the effort involved in each case. Replay attacks require slightly increased effort (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 client's 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. We suppose that replay attacks are less effective – though the experimental validation is lacking in the literature. This paper aims to assess this hypothesis.

3 Experimental setup

In the following we describe the ASV systems used in this study, the datasets, protocols and metrics, and then the implementation of each of the different spoofing attacks considered, including playback.

3.1 ASV systems

We assessed the impact of each spoofing attacks on six 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 [CSRS06], (iv) a GSL with factor analysis (FA) [FMS$^+$07], (v) a GMM-UBM system with factor analysis, and (vi) a state-of-the-art iVector system [DKD$^+$11].

The iVector system employs intersession compensation with probabilistic linear discriminant analysis (PLDA) [LFM$^+$12] with length normalisation [GREW11]. From here on in, it is referred to as the IV-PLDA system. The ASV systems were tested with and without normalisation. The IV-PLDA system used symmetric score normalisation (S-norm) as described in [Ken10], while the remaining systems utilised standard T-norm normalisation.

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 [BBF$^+$04]. Such an approach has been used successfully in many independent studies [MCGB01, FBK$^+$08].

All ASV systems were based on the LIA-SpkDet toolkit [BSM$^+$08] and the ALIZE library [BSFM04] and were directly derived from the work in [FMS$^+$07]. 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.

3.2 Datasets, protocols and metrics

All experiments reported below were performed using the male subsets of the standard 2005 and 2006 NIST Speaker Recognition Evaluation datasets (NIST'05 and NIST'06), distributed via the Linguistic Data Consortium (LDC). NIST'05 was used for optimising the ASV configurations whereas all results reported later relate to NIST'06, which was used for evaluation only.

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.

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

3.3 Spoofing attack

The setup of the four considered attacks is presented in the following, in the order of effort required – from zero-effort naïve impostors to high-effort speech synthesis.

The simplest, zero-effort attack consists in challenging the ASV system with the voice of a *naïve impostor*. This particular setup corresponds to the NIST baseline performance, which was assessed according to the protocol described in Section 3.2.

To emulate *replay attacks* at the sensor level we reproduce the distortions caused by a replay device and the effects introduced in typical acoustic conditions. We decided to use the speaker of a popular smartphone brand as the playback device (with the impulse

Norm	Attack	GMM	GSL	GSL-NAP	GSL-FA	FA	IV
No norm	Naïve impostor	9.08	7.89	6.35	6.08	5.60	3.04
	Replay	32.91	28.50	28.73	26.39	29.27	29.59
	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
With norm	Naïve impostor	8.63	8.13	6.31	5.72	5.61	2.98
	Replay	32.62	28.63	25.68	24.25	25.88	30.69
	Voice conversion	33.69	36.92	27.58	23.97	23.96	19.30
	Speech synthesis	27.29	15.04	13.78	11.91	16.22	10.82

Table 2: EER values for different ASV systems for various spoofing attacks, without and with score normalisation.

responses publicly available[2]) and an office room, as a likely environment for a spoofing attack. 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, was taken from the Aachen Impulse Response (AIR) database [JSV09].

Voice conversion was conducted with our implementation of the approach originally proposed in [MBC05]. 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[3] with a default configuration. We used standard speaker-independent models provided with the toolkit which were trained on the EMIME corpus [Wes10]. 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.

4 Results

Table 2 shows EER results for replay and other attacks against the six various ASV systems, with and without score normalisation. The results for naïve, zero-effort impostors correspond to baseline performance of the examined ASV systems and are in line with what can be expected from such systems in text-independent "one conversation" tasks – the IV-PLDA system performs best (EER of 2.98% with score normalisation), while the basic GMM-UBM yields the worst results.

For all other attacks presented in Table 2, all genuine client tests were unchanged, whereas impostor tests were replaced with spoofed accesses. All systems are shown to be severely

[2]http://www.aaronbrownsound.com/

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

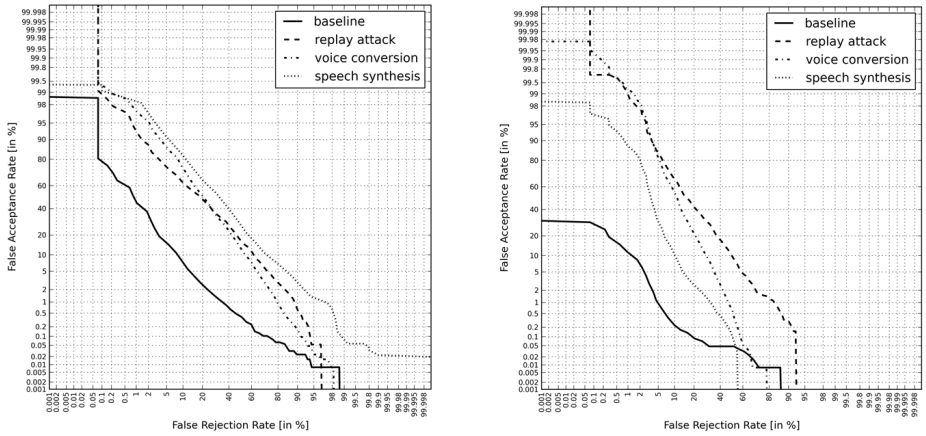

Figure 1: DET plots for GMM-UBM (left) and iVector-PLDA (right) systems.

sensitive to replay attacks – even for the most resistant (in terms of EER) GSL kernel system with factor analysis and with T-norm, the EER rose to 24%. The GMM-UBM and IV-PLDA systems yielded the worst results for replay – here the EER increased to more than 30%.

The impact of voice conversion, despite demanding considerably more effort to implement, causes a similar degradation in performance to that of replay attacks, with the exception of the IV-PLDA system which is more resistant to voice conversion than replay (19% EER vs. 31%, respectively). High-effort speech synthesis attacks proved even less effective – the EER for the best IV-PLDA system reached only 10.82%. These observations are also illustrated through detection error trade-off (DET) plots[4] in Fig. 1.

For all the ASV systems, test or score normalisation mostly helped to decrease EER values in the face of spoofing, e.g., for the factor analysis system the EER decreased from more than 30% to around 16%. In contrast, in some cases, e.g., for replay attacks and the IV-PLDA and GSL systems, the EER slightly increased after applying score normalisation.

The ambiguous impact of score normalisation is also visible in Table 3. It shows the EER and SFAR results for the simplest ASV system (GMM-UBM), the best system in the sense of baseline performance (IV-PLDA), and the system which showed the best robustness to replay attacks (GSL-FA). For calculating the SFAR result, the operating point was set to the baseline EER of the given system. Surprisingly, the IV-PLDA system showed the highest SFAR values for replay attacks – more than 80% of false acceptances, both with and without score normalisation. The shape of the DET plot towards the low false reject region in Fig. 2 confirms the high vulnerability of the IV-PLDA system to replay attacks compared to other systems.

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

ASV	Attack	EER (%)		SFAR (%)	
		no-norm	norm	no-norm	norm
GMM-UBM	Naïve impostor	9.08	8.63	9.08	8.63
	Replay	32.91	32.62	66.01	58.86
	Voice conversion	31.48	33.69	60.05	91.37
	Speech synthesis	39.90	27.29	87.14	71.83
IV-PLDA	Naïve impostor	3.04	2.98	3.04	2.98
	Replay	29.59	30.69	92.56	80.86
	Voice conversion	20.45	19.30	93.86	84.67
	Speech synthesis	10.92	10.82	43.78	30.00
GSL-FA	Naïve impostor	6.08	5.72	6.08	5.72
	Replay	26.39	24.25	60.91	60.49
	Voice conversion	30.23	23.97	89.68	73.16
	Speech synthesis	11.98	11.91	39.54	56.49

Table 3: Comparison of ASV performance in terms of EER and SFAR for the GMM-UBM, IV-PLDA and SGL-FA systems (without and with normalisation) for various spoofing attacks. For SFAR, FRR is set to the baseline EER.

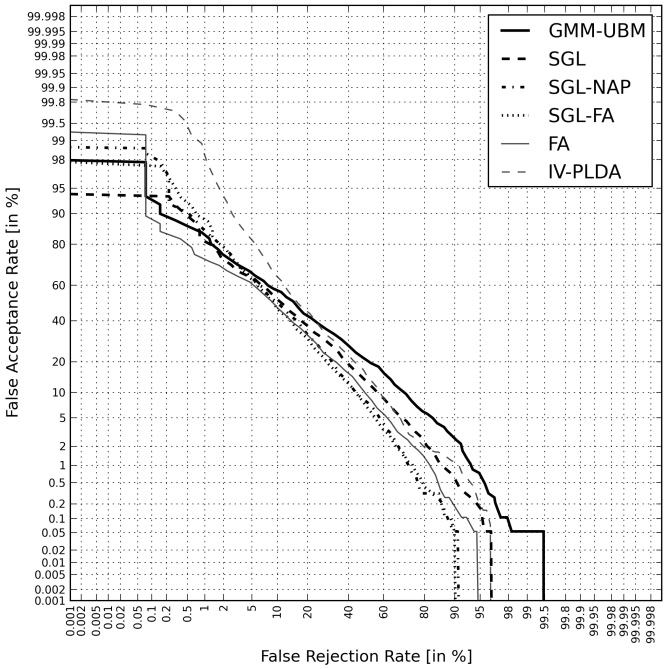

Figure 2: DET plots for replay attack challenging six various ASV systems, with test/score normalisation.

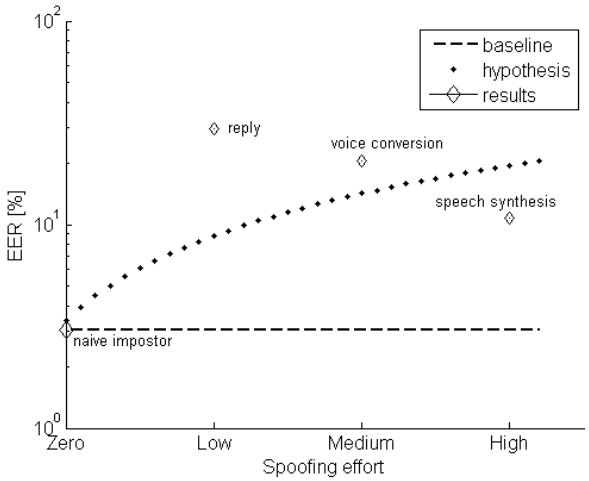

Figure 3: EER results for iVector-PLDA system against spoofing attacks with various effort level

5 Conclusions

This paper re-assesses the threat of replay attacks against automatic speaker verification (ASV) systems. The work was performed using simulated attacks and with large, standard NIST speaker recognition corpora and six ASV systems.

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 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

[AVE12] Federico Alegre, Ravichander Vipperla, and Nicholas Evans. Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial

signals. In *Proc. 13th Interspeech*, 2012.

[BBF+04] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 2004:430–451, January 2004.

[BMF06] Jean-François Bonastre, Driss Matrouf, and Corrine Fredouille. Transfer function-based voice transformation for speaker recognition. In *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, pages 1–6, 2006.

[BMF07] Jean-François Bonastre, Driss Matrouf, and Corinne Fredouille. Artificial impostor voice transformation effects on false acceptance rates. In *Proc. Interspeech*, pages 2053–2056, 2007.

[BSFM04] Jean-François Bonastre, Nicolas Scheffer, Corrine Fredouille, and Driss Matrouf. NIST'04 speaker recognition evaluation campaign: new LIA speaker detection platform based on ALIZE toolkit. In *NIST SRE'04*, 2004.

[BSM+08] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corrine Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoit Fauve, and John Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, volume 5, page 1, 2008.

[CSRS06] William M. Campbell, Douglas Sturim, Douglas A. Reynolds, and Alex 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)*, volume 1, page I, may 2006.

[DKD+11] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[EKY13] Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. Spoofing and countermeasures for automatic speaker verification. In *Proc. Interspeech 2013*, Lyon, France, 2013.

[EKY+14] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre, and Phillip L. De Leon. *Speaker recognition anti-spoofing*. Springer, 2014.

[FBK+08] Benoıt Fauve, Hervé Bredin, Walid Karam, Florian Verdet, Aurélien Mayoue, Gérard Chollet, Jean Hennebert, Richard Lewis, John Mason, Chafic 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*, pages 4137–4140. IEEE, 2008.

[FMS+07] Benoıt Fauve, Driss Matrouf, Nicolas Scheffer, Jean-François Bonastre, and John 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*, 15(7):1960–1968, 2007.

[FWAH08] Mireia Farrús, Michael Wagner, Jan Anguita, and Javier Hernando. How vulnerable are prosodic features to professional imitators? In *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2008.

[GREW11] Daniel Garcia-Romero and Carol Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *International Conference on Speech Communication and Technology*, pages 249–252, 2011.

[JSV09] Marco Jeub, Magnus Schäfer, and Peter Vary. A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms. In *Proceedings of the 16th International Conference on Digital Signal Processing*, DSP'09, pages 550–554, Piscataway, NJ, USA, 2009. IEEE Press.

[Ken10] Patrick Kenny. Bayesian Speaker Verification with Heavy-Tailed Priors. In *Odyssey*, page 14, 2010.

[LAPY10] Phillip L. De Leon, Vijendra Raj Apsingekar, Michael Pucher, and Junichi Yamagishi. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 1798 –1801, March 2010.

[LB99] Johan Lindberg and Mats Blomberg. Vulnerability in speaker verification - a study of technical impostor techniques. In *European Conference on Speech Communication and Technology*, pages 1211–1214, 1999.

[LFM+12] Peng Li, Yun Fu, Umar Mohammed, J.H. Elder, and S.J.D. Prince. Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):144–157, 2012.

[MBC05] Driss Matrouf, Jean-François Bonastre, and Jean-Pierre Costa. Effect of impostor speech transformation on automatic speaker recognition. *Biometrics on the Internet*, page 37, 2005.

[MCGB01] Ivan Magrin-Chagnolleau, Guillaume Gravier, and Raphaël Blouet. Overview of the 2000-2001 ELISA consortium research activities. In *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[MHTK99] Takashi Masuko, Takafumi Hitotsumatsu, Keiichi Tokuda, and Takao Kobayashi. On the Security of HMM-Based Speaker Verification Systems Against Imposture using Synthetic Speech. In *Proc. EUROSPEECH*, 1999.

[PAB+05] Patrick Perrot, Guido Aversano, Raphaël Blouet, Maurice Charbit, and Gérard Chollet. Voice Forgery Using ALISP : Indexation in a Client Memory. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, volume 1, pages 17 – 20, 2005.

[RM85] 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)*, pages 5–8, 1985.

[VL10] Jesus Villalba and Eduardo Lleida. Speaker verification performance degradation against spoofing and tampering attacks. In *FALA workshop*, pages 131–134, 2010.

[Wes10] Mirjam Wester. The EMIME bilingual database. Technical report, The University of Edinburgh, 2010.

[YKN+09] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai. Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. *IEEE transactions on Audio, Speech & Language Processing*, 17(1):66–83, 2009.

[YNZ+09] Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. Robust Speaker adaptive HMM based Text-to-Speech Synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 17(6):1208–1230, 2009.