Spoofing and countermeasures for automatic speaker verification

Nicholas Evans[1]*, Tomi Kinnunen*[2] *and Junichi Yamagishi*[3,4]

[1]EURECOM, Sophia Antipolis, France, [2]University of Eastern Finland, Finland
[3]University of Edinburgh, UK, [4]National Institute of Informatics, Japan
evans@eurecom.fr, tomi.kinnunen@uef.fi, jyamagis@inf.ed.ac.uk

Abstract

It is widely acknowledged that most biometric systems are vulnerable to spoofing, also known as imposture. While vulnerabilities and countermeasures for other biometric modalities have been widely studied, e.g. face verification, speaker verification systems remain vulnerable. This paper describes some specific vulnerabilities studied in the literature and presents a brief survey of recent work to develop spoofing countermeasures. The paper concludes with a discussion on the need for standard datasets, metrics and formal evaluations which are needed to assess vulnerabilities to spoofing in realistic scenarios without prior knowledge.

Index Terms: spoofing, imposture, automatic speaker verification

1. Introduction

Over the last decade biometrics technologies have revolutionised our approach to personal identification and have come to play an essential role in safeguarding personal, national and global security. It is widely acknowledged, however, that biometric systems can be fooled or 'spoofed' [1].

Efforts to develop spoofing countermeasures are under way across the various biometrics communities[1]. Progress in the case of automatic speaker verification (ASV) is, however, less advanced than for some other biometric modalities. Furthermore, since ASV is commonly used in telephony, or other unattended, distributed scenarios without human supervision or face-to-face contact, speech is arguably more prone to malicious interference or manipulation than other biometric signals.

Previous efforts to develop countermeasures for ASV [2, 3, 4, 5] generally exploit prior knowledge of specific spoofing attacks and usually focus on text-independent ASV. The use of prior knowledge is clearly unrepresentative of the practical scenario where the nature of the attack can never be known. There is thus a need to collect public datasets of licit and spoofed speaker verification transactions to facilitate independent efforts in spoofing assessment and the development of countermeasures which are less dependent on prior knowledge. Ultimately, this initiative will require the expertise of different speech and language processing communities, e.g. those in voice conversion and speech synthesis, in addition to ASV.

The Interspeech 2013 special session in Spoofing and Countermeasures for Automatic Speaker Verification was organised by the authors of this paper to encourage the discussion and collaboration needed to organise the collection of standard datasets and the definition of metrics and evaluation protocols for future research in spoofing and countermeasures for ASV. This paper aims to provide the starting point for such an initiative. It describes a selection of vulnerabilities studied previously, presents a brief survey of recent work to develop spoofing countermeasures and discusses current approaches to evaluation.

The remainder of this paper is organised as follows. Section 2 describes state-of-the-art approaches to speaker verification and accounts for their vulnerability to spoofing. Past work to assess those vulnerabilities is presented in Section 3 with an account of related efforts to develop suitable countermeasures. We discuss different approaches to assessment and the need to develop standard databases, metrics and assessment protocols in Section 4. Conclusions are presented in Section 5.

2. Automatic speaker verification

The paper focuses on text-independent ASV. In this section we describe state-of-the-art approaches and their vulnerability to spoofing.

2.1. Feature extraction

Speech production is a highly non-stationary process. Since the acoustic characteristics change continuously over time, features are commonly extracted from short-term segments (frames) of 20 to 30 msec in duration. Typical feature extractors find a low-dimensional parametrisation for the short-term power spectrum of speech, e.g. mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs) and perceptual linear prediction (PLP) features. These features are commonly appended with their time derivatives (delta and double delta features) and are generally normalized, e.g. through mean removal or short-term Gaussianization [6]. Further details can be found in, e.g. [7]. As discussed later, the literature shows that speech signals with short-term instantaneous spectral representations indicative of other speakers can be readily synthesized.

2.2. Speaker and session modelling

Approaches to text-independent ASV generally focus on modelling the long-term distribution of spectral vectors, for which Gaussian mixture models (GMMs) [8, 9] have become the de facto standard. The speaker verification systems of the 1990s and early 2000s used either maximum likelihood (ML) [8] or maximum a posteriori (MAP) [9] criteria to train speaker-dependent GMMs. In the latter case, speaker-dependent GMMs are obtained from the adaptation of pre-trained universal background models (UBMs). Adapted GMM mean supervectors obtained in this way were later successfully combined with support vector machines (SVMs) [10]. This involved the development of trainable intersession variability compensation techniques such as nuisance attribute projection (NAP) [11, 12] and

[1]http://www.tabularasa-euproject.org/

within-class covariance normalization (WCCN) [13].

Parallel to these developments in SVM-based discriminative speaker modelling, mathematically rigorous, generative factor analysis models were pioneered in [14, 15, 16]. The so-called joint factor analysis (JFA) technique [14] achieved state-of-the-art accuracy using separate mechanisms to model speaker and session variability. This technique was further simplified in the total variability model [17], commonly referred to as the i-vector framework which, in contrast to JFA, does not differentiate between speaker and session subspace models. i-vectors are typically of 200 to 600 dimensions and contain both speaker and channel variations; unwanted variability is handled in back-end classification with, e.g. probabilistic linear discriminant analysis (PLDA) [18]. Practice has also shown the benefit of normalizing i-vectors to have unit norm (i.e. so they lie on a hypersphere) is helpful [19].

Even if there is evidence that the more sophisticated approaches to ASV are more resilient to spoofing, all have their roots in the standard GMM. Assuming independent observations, none utilises time sequence information, a key characteristic of speech which might otherwise protect systems from spoofing.

3. Spoofing and countermeasures

Spoofing attacks are performed on a biometric system at the sensor or acquisition level to bias score distributions toward those of genuine clients, thus provoking increases in the false acceptance rate (FAR). This section reviews past work to evaluate vulnerabilities and to develop spoofing countermeasures. We consider impersonation, replay, speech synthesis and voice conversion. We stress that in all cases we retain the emphasis on text-independent ASV, i.e. we do not consider text-dependent nor challenge-response countermeasures.

3.1. Impersonation

Impersonation refers to spoofing attacks with human-altered voices and is one of the most obvious forms of spoofing and earliest studied.

3.1.1. Spoofing

The work in [2] showed that non-professional impersonators can readily adapt their voice to overcome ASV, but only when their natural voice is already similar to that of the target. Further work in [20] showed that impersonation increased FAR rates from close to 0% to between 10% and 60%, but no significant difference in vulnerability to non-professional or professional impersonators. Characteristic to these studies is the use of relatively few speakers.

3.1.2. Countermeasures

None of the above studies investigated countermeasures against impersonation. Impersonation involves mostly the mimicking of prosodic or stylistic cues rather than those aspects more related to the vocal tract. Impersonation is therefore considered more effective in fooling human listeners than a genuine threat to today's state-of-the-art ASV systems [4].

3.2. Replay

Replay attacks [21] involve the presentation of speech samples captured from a genuine client in the form of continuous speech recordings, or samples resulting from the concatenation of shorter segments [21].

3.2.1. Spoofing

While some form of text-dependent or challenge-response countermeasure is usually used to prevent replay-attacks, text-independent solution have also been investigated. Work in [22] investigated vulnerabilities to the replaying of far-field recorded speech. Using a baseline ASV system based on JFA, their work showed an increase in the equal error rate (EER) of 1% to almost 70% when imposter accesses were replaced by replayed spoof attacks.

3.2.2. Countermeasures

The same authors showed that it is possible to detect such spoofing attacks by measuring the channel differences caused by far-field recording [3]. While they show spoof detection error rates of less than 10% it is feasible that today's state-of-the-art approaches to channel compensation will leave some systems more vulnerable to replay attacks.

3.3. Speech synthesis

There are two major approaches to speech synthesis: unit selection and statistical parametric approaches. The unit selection approach generally requires large amounts of speaker-specific data with carefully prepared transcripts in order to construct speech models. In contrast, state-of-the-art hidden Markov model (HMM)-based speech synthesizers [23] can learn speech models from relatively little speaker-specific data and the adaptation of background models derived from other speakers. There is a considerable volume of research in the literature which has demonstrated the vulnerability of ASV to synthetic voices.

3.3.1. Spoofing

ASV vulnerabilities to synthetic speech were first demonstrated over a decade ago [24] using an HMM-based, text-prompted ASV system [25] and an HMM-based synthesizer where acoustic models were adapted to specific human speakers [26, 27]. The ASV system scored feature vectors against speaker and background models composed of concatenated phoneme models. When tested with human speech the ASV system achieved an FAR of 0% and a false rejection rate (FRR) of 7%. When subjected to spoofing attacks with synthetic speech, the FAR increased to over 70%, however this work involved only 20 speakers.

Larger scale experiments using the Wall Street Journal corpus containing in the order of 300 speakers and two different ASV systems (GMM-UBM and SVM using Gaussian supervectors) was reported in [28]. Using a state-of-the-art HMM-based speech synthesiser, the FAR was shown to rise to 91%. Spoofing experiments using HMM-based synthetic speech against a forensics speaker verification tool *BATVOX* was reported in [29] with similar findings. Today's state-of-the-art speech synthesizers thus present a genuine threat to ASV.

3.3.2. Countermeasures

Only a small number of attempts to discriminate synthetic speech from natural speech have been investigated and there is currently no general solution which is independent from specific speech synthesis methods. Previous work has demon-

strated the successful detection of synthetic speech based on prior knowledge of the acoustic differences of specific speech synthesizers, such as the dynamic ranges of spectral parameters at the utterance level [5] and variance of higher order parts of mel-cepstral coefficients [30].

There are some attempts which focus on acoustic differences between vocoders and natural speech. Since the human auditory system is known to be relatively insensitive to phase [31], vocoders are typically based on a minimum-phase vocal tract model. This simplification leads to differences in the phase spectra between human and synthetic speech, differences which can be utilised for discrimination [28, 32].

Other approaches to synthetic speech detection use F0 statistics [33, 34], based on the difficulty in reliable prosody modelling in both unit selection and statistical parametric speech synthesis. F0 patterns generated for the statistical parametric speech synthesis approach tend to be over-smoothed and the unit selection approach frequently exhibits 'F0 jumps' at concatenation points of speech units.

3.4. Voice conversion

Voice conversion is a sub-domain of voice transformation [35] which aims to convert one speaker's voice towards that of another [35]. The field has attracted increasing interest in the context of ASV vulnerabilities for over a decade [36].

3.4.1. Spoofing

When applied to spoofing, the aim with voice conversion is to synthesize a new speech signal such that extracted ASV features are close in some sense to the target speaker. Some of the first work relevant to text-independent ASV spoofing includes that in [4, 37]. The work in [4] showed that a baseline EER increased from 16% to 26% as a result of voice conversion which also converted prosodic aspects not modelled in typical ASV systems. The work in [37] investigated the probabilistic mapping of a speaker's vocal tract information towards that of another, target speaker using a pair of tied speaker models, one of ASV features and another of filtering coefficients. This work targeted the conversion of spectral-slope parameters. The work showed that a baseline EER of 10% increased to over 60% when all impostor test samples were replaced with converted voice. In addition, signals subjected to voice conversion did not exhibit any perceivable artefacts indicative of manipulation.

The work in [38] investigated ASV vulnerabilities using a popular approach to voice conversion [39] based on joint-density GMMs, which requires a parallel training corpus for both source and target speakers. Even if converted speech is usually detected by human listeners, experiments involving five different ASV systems showed universal susceptibility to spoofing. The FAR of the most robust, JFA system increased from 3% to over 17%.

Other work relevant to voice conversion includes attacks referred to as artificial signals. It was noted in [40] that certain short intervals of converted speech yield extremely high scores or likelihoods. Such intervals are not representative of intelligible speech but they are nonetheless effective in overcoming typical ASV systems which lack any form of speech quality assessment. The work in [40] showed that artificial signals optimised with a genetic algorithm provoke increases in the EER from 10% to almost 80% for a GMM-UBM system and from 5% to almost 65% for a factor analysis (FA) system.

3.4.2. Countermeasures

Some of the first work to detect converted voice draws on related work in synthetic speech detection [41]. While the proposed cos-phase and modied group delay function (MGDF) phase countermeasures proposed in [32] are effective in detecting synthetic speech, they are unlikely to detect converted voice with real-speech phase [37].

Two approaches to artificial signal detection are reported in [42]. Experimental work shows that supervector-based SVM classifiers are naturally robust to such attacks whereas all spoofing attacks can be detected using an utterance-level variability feature which detects the absence of natural, dynamic variability characteristic of genuine speech. An alternative approach based on voice quality analysis is less dependent on explicit knowledge of the attack but less effective in detecting attacks.

A related approach to detect converted voice is proposed in [43]. Probabilistic mappings between source and target speaker models are shown to yield converted speech with less short-term variability than genuine speech. The thresholded, average pair-wise distance between consecutive feature vectors is used to detect converted voice with an EER of under 3%.

4. Discussion

In the following we discuss current approaches to evaluation and some weaknesses in research and evaluation methodology.

4.1. Protocols and metrics

While countermeasures can be integrated into existing ASV systems, they are most often implemented as independent modules which allow for the explicit detection of spoofing attacks. The most common approach in this case is to concatenate the two classifiers in series.

The assessment of countermeasure performance on its own is relatively straightforward; results are readily analysed with standard detection error trade-off (DET) profiles and related metrics. It is often of interest, however, that the assessment reflects their impact on ASV performance. Assessment is then non-trivial and calls for the joint optimisation of combined classifiers. Results furthermore reflect the performance of non-standard ASV systems. As reflected in Section 3, there are currently no standard protocols, metrics or ASV systems which might otherwise be used to conduct fair evaluations with comparable results. There is a thus a need to define such standards in the future.

Candidate standards are being drafted within the scope of the EU FP7 TABULA RASA project. Here, independent countermeasures preceding biometric verification are optimised at three different operating points where thresholds are set to obtain FARs (the probability of labelling a genuine access as a spoofing attack) of either 1, 5 or 10%. Samples labelled as genuine accesses are then passed to the verification system[2]. Performance is assessed using four different DET profiles[3], examples of which are illustrated in Figure 1. The four profiles illustrate performance of the baseline system with naïve impostors, the baseline system with active countermeasures, the baseline system where all impostor accesses are replaced with spoofing

[2] In practice samples labelled as spoofing attacks cannot be fully discarded since so doing would unduly influence false reject and false acceptance rates calculated as a percentage of all accesses.

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

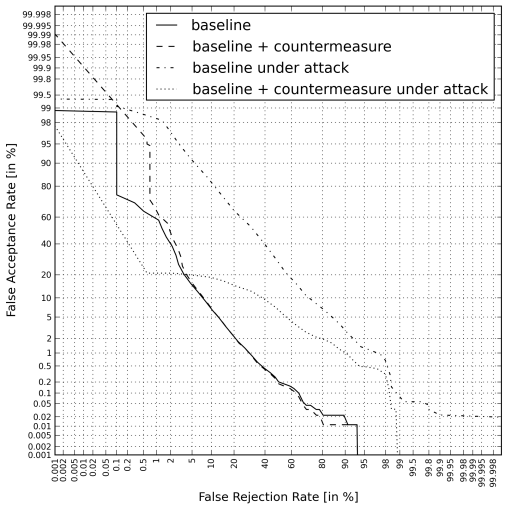

Figure 1: An example of four DET profiles which can be used to analyse vulnerabilities to spoofing and countermeasure performance. Results correspond to spoofing attacks using synthetic speech and a standard GMM-UBM classifier assessed on the male subset of the NIST'06 SRE dataset.

attacks and, finally, the baseline system with spoofing attacks and active countermeasures.

Consideration of all four profiles is needed to gauge the impact of countermeasure performance on licit transactions (any deterioration in false rejection – difference between 1^{st} and 2^{nd} profiles) and improved robustness to spoofing (improvements in false acceptance – difference between 3^{rd} and 4^{th} profiles). While the interpretation of such profiles is trivial, different plots are obtained for each countermeasure operating point. Further work is required to design intuitive, universal metrics which represent the performance of spoofing countermeasures when combined with ASV.

4.2. Datasets

While some work has shown the potential for detecting spoofing without prior knowledge or training data indicative of a specific attack [32], all previous work is based on some implicit prior knowledge, i.e. the nature of the spoofing attack and/or the targeted ASV system is known. While training and evaluation data with known spoofing attacks might be useful to develop and optimise appropriate countermeasures, the precise nature of spoofing attacks can never be known in practice. Estimates of countermeasure performance so obtained should thus be considered at best optimistic. Furthermore, some of the past work was also conducted under matched conditions, i.e. data used to learn target models and that used to effect spoofing were collected in the same or similar acoustic environment and over the same or similar channel. The performance of spoofing countermeasures when subjected to realistic session variability is then unknown.

While much of the past work already uses standard datasets, e.g. NIST SRE data, spoofed samples are obtained by treating them with non-standard algorithms. Standard datasets containing both licit transactions and spoofed speech from a multitude of different spoofing algorithms and with realistic session variability are therefore needed to reduce the use of prior knowledge, to improve the comparability of different countermeasures and their performance against varied spoofing attacks. Collaboration with colleagues in other speech and language processing communities, e.g. voice conversion and speech synthesis, will help to assess vulnerabilities to state-of-the art spoofing attacks and also to assess countermeasures when details of the spoofing attacks are unknown. The detection of spoofing will then be considerably more challenging but more reflective of practical use cases.

5. Conclusions

This paper gives an overview of recent research in spoofing and countermeasures for ASV. While it is clear that ASV systems can be vulnerable to spoofing, most vulnerabilities discussed in this paper involve relatively high-cost, high-technology attacks. Furthermore, countermeasures, some of them relatively trivial, have the potential to detect spoofing attacks with manageable impacts on system usability. Further work should analyse the potential for spoofing through risk assessment and address some weaknesses in the current research methodology.

The Interspeech 2013 Special Session on Spoofing and Countermeasures was organised by the authors of this paper to promote the consideration of spoofing, to encourage the development of countermeasures and to form a new community of researchers to organise the next steps towards formal evaluations. Closer collaboration is needed to collect standard datasets containing both genuine and spoofed speech and thus to facilitate the development of universal, robust countermeasures capable of detecting unforeseen spoofing attacks.

6. Acknowledgements

This work was partially supported by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU) (grant agreement number 257289), by the Academy of Finland (project no. 253120) and by EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

7. References

[1] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[2] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*. IEEE, 2004, pp. 145–148.

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

[4] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP: indexation in a client memory," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 17–20.

[5] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eurospeech*, 2001.

[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of Odyssey 2001: The Speaker and Language Recognition Workshop*, Crete, Greece, June 2001, pp. 213–218.

[7] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[8] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[10] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[11] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proceedings of ICASSP 2005*, Philadelphia, USA, March 2005, pp. 629–632.

[12] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1979–1986, September 2007.

[13] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. of the Int. Conf. on Spoken Language Processing*, September 2006, pp. 1471–1474.

[14] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," technical report CRIM-06/08-14, Montreal, CRIM, 2006.

[15] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[16] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

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

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

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[20] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the yoho speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2005, pp. 907–907.

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

[22] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 10 workshop*, 2010, pp. 131–134.

[23] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[24] 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.

[25] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Commun.*, vol. 17, no. 1-2, pp. 109–116, Aug. 1995.

[26] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP*, 1996.

[27] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP*, 1997.

[28] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[29] G. Galou, "Synthetic voice forgery in the forensic context: a short tutorial," in *Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG)*, Sep. 2011, pp. 1–3.

[30] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, 29 2010-Dec. 3, pp. 309–312.

[31] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice-Hall, Inc., 2002.

[32] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Interspeech 2012*, 2012.

[33] A. Ogihara, H. Unno, and A. Shiozakai, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 1, pp. 280–286, jan 2005.

[34] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Interspeech*, Portland, Oregon, USA, Sep. 2012.

[35] Y. Stylianou, "Voice transformation: a survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3585–3588.

[36] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 837–840.

[37] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[38] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4401–4404.

[39] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 285–288.

[40] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *European Conference on Signal Processing (EUSIPCO), 2012 EURASIP Conference on*. EURASIP, 2012.

[41] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, USA, 2011, pp. 4844–4847.

[42] F. Alegre, R. Vipperla, N. Evans *et al.*, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, 2012.

[43] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speker verification from voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013.