EVASION AND OBFUSCATION IN SPEAKER RECOGNITION SURVEILLANCE AND FORENSICS

Federico Alegre[1]*, Giovanni Soldi*[1]*, Nicholas Evans*[1]*, Benoit Fauve*[2] *and Jasmin Liu*[2]

[1]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
[2]ValidSoft Ltd, London, UK

{alegre,soldi,evans}@eurecom.fr, {benoit.fauve,jasmin.liu}@validsoft.com

ABSTRACT

This paper presents the first investigation of evasion and obfuscation in the context of speaker recognition surveillance and forensics. In contrast to spoofing, which aims to provoke false acceptances in authentication applications, evasion and obfuscation target detection and recognition modules in order to provoke missed detections. The paper presents our analysis of each vulnerability and the potential for countermeasures using standard NIST datasets and protocols and six different speaker recognition systems (from a standard GMM-UBM system to a state-of-the-art i-vector system). Results show that all systems are vulnerable to both evasion and obfuscation attacks and that a new generalised countermeasure shows promising detection performance. While all evasion attacks and almost all obfuscation attacks are detected in the case of this particular setup, the work nonetheless highlights the need for further research.

Index Terms— evasion, obfuscation, speaker recognition, surveillance, forensics, biometrics, spoofing

1. INTRODUCTION

It is now well known that most biometric systems are vulnerable to some form of subversion [1]. Subversion aims to provoke a recognition error, either a false acceptance in the case of authentication applications, or a missed detection in the case of surveillance and forensics.

Authentication applications include access control and general security scenarios. They typically involve the identification or verification of a pre-enrolled individual seeking access to protected resources. The threat in this scenario involves spoofing, where an impostor impersonates the biometric traits of an enrolled individual in order to provoke a false acceptance.

Surveillance and forensic investigation applications include detection tasks, e.g. the detection of known speakers in intercepted telephone conversations or other audio evidence. The threat in this scenario involves evasion and obfuscation, whereby a person of interest might seek to provoke a missed detection.

There is arguably a third form of subversion, related more closely to traditional forensics, for example the analysis of DNA, fingerprints, hair samples, voice recordings etc. Here, biometric evidence can be manipulated, not only to evade reliable detection, but also so that they indicate the identity of another, specific person, i.e. to implicate another individual through the fabrication of false evidence.

While spoofing research in automatic speaker verification (ASV) is only just beginning to gather pace [2], there is almost no work in the literature related to either evasion or obfuscation [3]. Reliable recognition performance is essential whatever the application. Spoofing can result in the granting of access to critical resources to persons of ill intent, whereas evasion and obfuscation can encumber or jeopardise criminal convictions. It is thus essential that studies of evasion and obfuscation are made in parallel with, and to accelerate the design of new approaches to detect manipulated evidence and to ensure the reliability of ASV in surveillance and forensic applications.

This paper reports our study of evasion and obfuscation in the context of ASV. The potential to provoke missed detections is assessed using six different ASV systems, from a standard GMM-UBM system to a state-of-the art i-vector system. New to this contribution is the classification and study of independent evasion and obfuscation attacks. Also reported is a new countermeasure which aims to identify attempts to evade and obfuscate detection.

2. EVASION AND OBFUSCATION

Both evasion and obfuscation refer to the intentional manipulation of a biometric signal in order to provoke missed detections. While the notion of obfuscation is now widely understood [4] we see fundamental differences between evasion and obfuscation; while the end result is the same, the attacks target two distinctly different components of a typical biometric system.

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).

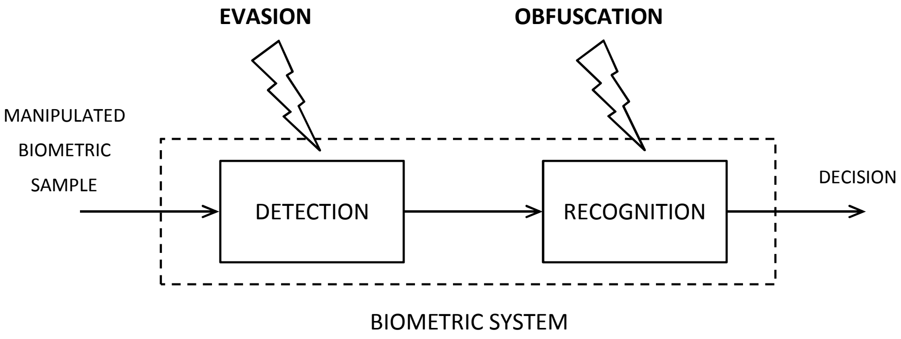

Fig. 1. Scenarios for evasion and obfuscation.

Evasion and obfuscation attacks are illustrated in Figure 1. It shows the two critical elements in a standard biometric system, namely the detection and recognition modules, which may be vulnerable to evasion and obfuscation respectively.

2.1. Evasion

The detection module aims to identify those components or intervals of the input signal which are of interest to the recognition module, i.e. typically the components containing a face, a fingerprint, or the intervals containing speech. Evasion attacks can be applied here to prevent such components from being identified. Consequently, the recognition module will never receive a valid biometric sample.

In terms of ASV, the biometric detector is commonly referred to as either speech activity detection (SAD) or voice activity detection (VAD). Three predominant forms are used in practice but, be they energy-based, model-based of phoneme-based detectors, the goal is common to all, namely to identify frames in the input signal which contain useful speech.

While model-based and phoneme-based approaches are more complex and conceivably more robust to evasion, energy-based approaches can be overcome with relative ease. Since they generally rely on relatively clean signal-to-noise ratios, a simple attack might involve the filling of non-speech periods with a signal whose energy is higher than that of the speech. As a result, only non-informative intervals which do not contain speech will then be passed to the recognition module.

Energy-based SADs are still well accepted in the literature and, mostly for reasons of computational simplicity, they might be preferred in practice. There is also some recent evidence [5] which suggests that energy-based SAD can be more effective than alternative model and phoneme-based SAD in a variety of noise conditions.

2.2. Obfuscation

Assuming that useful speech does reach the recognition stage, then here there is potential for the speech signal to be manipulated in order to interfere with the decision and once again provoke a recognition error. In line with the definition for fingerprint recognition [4] we refer to speech obfuscation as the intentional manipulation of an utterance in order to provoke missed detections.

In this context, obfuscation can be seen as a sub-domain of voice disguise, which considers both intentional and non-intentional speech alterations [6]. Other approaches might include automatic manipulations such as voice transformation and voice conversion [7], pitch modification (e.g. falsetto), whispering, glottal fry, pinched nostril speech, bite blocking, a hand over the mouth, imitation and other mechanical/prosody alterations.

There is very little work in the literature relating to obfuscation, despite convincing arguments supporting the potential. The work in [8, 9, 10, 11] investigated the effect of intentional voice modifications or disguise and found in all cases that missed detection rates increase. Automatic approaches to voice transformation reported in [3, 12] are also shown to overcome identification and verification systems, though most of this work involves the use of non-standard, small datasets. The first work to detect disguised voice is reported in [13]. While performed using the standard TIMIT database and while promising detection rates are reported, the work does not consider impacts on ASV performance.

This paper presents the first assessment of evasion and obfuscation under controlled conditions using large-scale, standard NIST databases and state-of-the-art approaches to obfuscation which have already been shown to overcome ASV through spoofing. We also present a new approach to evasion and obfuscation detection and analyse its impact on ASV performance.

3. EVALUATION

This section presents our work to assess the vulnerability of automatic speaker verification (ASV) to evasion and obfuscation. We describe the different ASV systems, datasets and protocols used in this work, the particular approaches to evasion and obfuscation, and experimental results.

We stress that the full consideration of every possible threat is beyond the scope of this contribution. Clearly this work is only a start to broader research which will require greater attention in the future.

3.1. ASV systems: speech detection and modelling

We assessed the impact of evasion and obfuscation on six different ASV systems: (i) a standard GMM-UBM system; (ii) a GMM-UBM system with factor analysis (FA) channel compensation; (iii-v) three different GMM supervector linear kernel (GSL) systems, and (vi) a state-of-the-art i-vector system.

The FA system is based on the approach described in [14]. The standard GSL system uses a support vector machine classifier which is applied to supervectors obtained with the GMM-UBM system. The second GSL system is enhanced with nuisance attribute projection [15] whereas the third uses FA supervectors (GSL-FA) [16]. The i-vector system [17] employs intersession compensation with probabilistic linear discriminant analysis (PLDA) [18] with length normalisation [19]. From here on in it is referred to as IV-PLDA.

All ASV systems use 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 [20]. As mentioned in Section 2.1, we note that, on similar data to that used here, such an approach performs well compared to alternatives for different types of noisy environments, e.g. model-based and phone-based [5], even if the consideration here includes predominantly low noise condition.

All ASV systems are based on the LIA-SpkDet toolkit [21] and the ALIZE library [22] and are directly derived from the work in [16]. They furthermore use a common UBM with 1024 Gaussian components, the speech activity detector already mentioned and feature parametrisation: linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy. Full details of all systems can be traced through [23].

3.2. Datasets and protocols

All development was performed using the male subset of the 2005 NIST Speaker Recognition Evaluation dataset (NIST‘05) whereas the male subset of the NIST‘06 dataset was used for evaluation. Only evaluation results are reported in this paper. The NIST‘04 or NIST‘08 datasets are used as background data, depending on whether the data is used for ASV or evasion and obfuscation respectively.

To assess the potential impact of evasion and obfuscation, true-client tests are replaced with alternative speech data which aims to either evade or obfuscate reliable recognition. Any number of different approaches may be used. The specific approaches chosen in each case are described in the next sections. The only difference between their use in the study of evasion and obfuscation instead of spoofing involve their application to client trials (instead of impostor trials) to provoke missed detections (instead of false accepts).

3.3. Evasion with white noise

The scale of the evasion threat will naturally depend on the specific approach to SAD. In the following, we assume the use of a simple, energy-based approach and report illustrative examples with a equally straightforward, targeted attack in order to demonstrate the concept.

The input speech signal is first processed offline to identify low-energy, neighbouring intervals of non-speech. The average energy level of the intervals containing speech is then estimated and the non-speech intervals alone are filled with higher-energy white noise. While the resulting signal is perceptually challenging, and with the exception of some masking effects, the speech remains entirely intelligible.

As described in Section 3.1, only the higher-energy components of the input signal are retained after SAD. Such a trivial attack thus succeeds in ensuring that very little, if any useful clean speech is passed to the speaker recognition system which instead receives only intervals of non-informative noise.

3.4. Obfuscation by voice conversion

The approach to obfuscation used in this work is based on voice conversion. It is applied here according to the Gaussian dependent filtering (GDF) approach proposed in [24]. The GDF approach converts the speech of an original speaker $y(n)$ towards that of a target speaker $x(n)$ in the spectral domain according to:

$$Y'(f) = \frac{|H_{\mathrm{x}}(f)|}{|H_{\mathrm{y}}(f)|} Y(f) \tag{1}$$

where $|H_{\mathrm{y}}(f)|$ and $|H_{\mathrm{x}}(f)|$ are the vocal tract transfer functions of the original and target speakers respectively and where $Y(f)$ and $Y'(f)$ are the Fourier domain representations of $y(n)$ and $y'(n)$, the conversion result.

$H_{\mathrm{x}}(f)$ 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 is utilised for the calculation of *a posteriori* probabilities. The second, denoted as the filtering (fil) model, is a tied model of linear predictive cepstral coding (LPCC) coefficients from which $H_{\mathrm{y}}(f)$ is derived. LPCC filter parameters are estimated according to:

$$x_{\mathrm{fil}} = \sum_{i=1}^{M} p(g^i_{\mathrm{asr}} | y_{\mathrm{asr}}) \mu^i_{\mathrm{fil}} \tag{2}$$

where $p(g^i_{\mathrm{asr}} | y_{\mathrm{asr}})$ is the *a posteriori* probability of Gaussian component g^i_{asr} given the frame y_{asr} and μ^i_{fil} is the mean of

	EER (%)				minDCF $\times 100$			
System	ASV	Evasion+ASV	Obf.+ASV	Obf.+ASV+CM	ASV	Evasion+ASV	Obf.+ASV	Obf.+ASV+CM
GMM-UBM	8.7	19.4	47.7	4.1	4.16	10.28	10.15	0.47
GSL	8.0	55.1	32.3	3.5	3.38	10.02	10.07	0.47
GSL-NAP	6.8	53.4	31.5	3.4	2.53	10.02	9.08	0.46
GSL-FA	6.4	54.7	29.1	3.1	2.34	10.00	8.68	0.46
FA	5.6	20.6	41.9	3.9	2.33	10.06	9.98	0.47
IV-PLDA	3.0	24.3	20.0	2.9	1.15	10.06	9.67	0.47

Table 1. ASV performance without speech alteration (baseline), with evasion with noise and obfuscation through voice conversion and also detection of converted voices. Results shown in terms of EER and minDCF $\times 100$. Note: detection of evasion is not included in the table since both EERs and minDCF are 0 in all the cases.

component g^i_{fil} which is tied to g^i_{asr}. $H_{\mathrm{x}}(f)$ 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 [24, 25, 26].

In order to simulate obfuscation, voice conversion is applied to all true-client test utterances. To increase the chances of provoking missed detections we further convert each utterance towards the most dissimilar speaker among a selection of 10 randomly chosen subjects (that for which the likelihood score from a conventional trial is the lowest).

3.5. Results

Baseline ASV results are presented together with those for evasion and obfuscation in Table 1. Results are illustrated in terms of the equal error rate (EER) and the minimum decision cost function (minDCF). We discuss only the former in the following.

Table 1 shows that, for GSL-based systems under evasion, the EER increases from in the order of 7% to over 50%. On the other hand, the baseline EERs for the GMM-UBM, FA and IV-PLDA systems increase from between 3% and 9% to between 19% and 24%. Perhaps surprisingly, the simplest GMM-UBM system is seemingly the most robust. This observation is even more surprising given that almost no speech is treated by the speaker recognition system. The variation in performance is accounted for by differing levels of overlap between the score distributions with and without evasion.

Table 1 also illustrates the effect of obfuscation. Results show that now it is the GMM-UBM system which is the most vulnerable; the EER increases from 9% to 48%. The FA and three GSL-based systems also show high levels of vulnerability (EERs between 29% and 42%), whereas the IV-PLDA system is the most robust; the EER increases from 3% to 20%. Even so, this equates to a relative increase of over 550% and shows that all ASV systems are vulnerable to obfuscation.

Figure 2 shows a histogram of scores for impostor trials (left-most distribution) and target trials (right-most). Also illustrated is the score distribution for obfuscation tests which shows how voice conversion is effective in decreasing the likelihood scores for target tests; the degree of overlap with the impostor distribution is higher than for the target distribution thus accounting for the increase in EER.

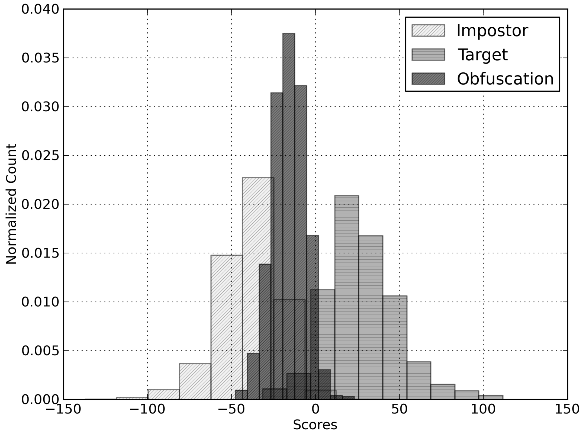

Fig. 2. IV-PLDA score distributions for impostor (left-most), target (right-most) and obfuscation trials with voice conversion.

Detection error trade-off (DET) profiles[1] for the IV-PLDA system are illustrated in Figure 3. Profiles for the baseline and obfuscation show that the system is vulnerable across the full range of operating points.

4. DETECTION

Various different approaches to detect manipulated speech signals have been reported in the literature. All involve the study of spoofing and the detection of processing artifacts indicative of manipulation, e.g. the absence of natural-speech phase [27] and reduced short-term dynamic variability [28].

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

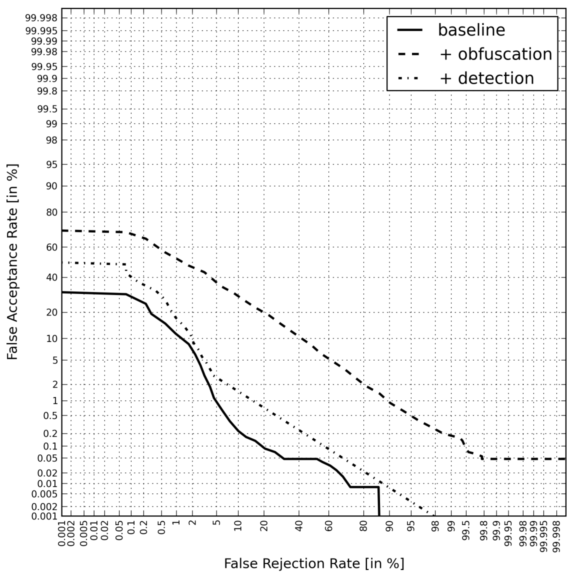

Fig. 3. DET profiles illustrating IV-PLDA performance for the baseline, the baseline with obfuscation attacks and then with integrated detection.

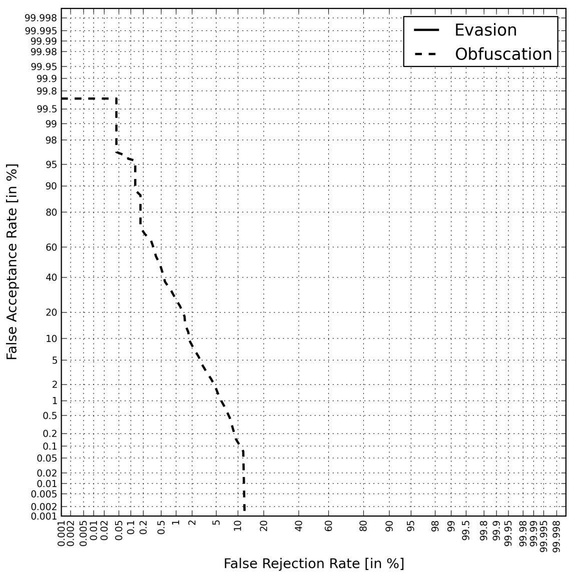

Fig. 4. A DET profile illustrating detection performance independently from ASV. The profile for evasion is not visible since the EER is 0%.

These approaches are, however, dependent on the specific approach to spoofing and thus have limited practical application. The work in [23] presented the first generalised solution with the potential to detect previously unseen approaches to manipulation. A new, one-class classification approach learnt using only genuine speech is used to detect the absence of natural spectro-temporal variability through the so-called local binary pattern (LBP) analysis of speech spectrograms. With improved generalisation, this approach to detection has greater practical application and is thus the approach adopted here as a means of detecting both evasion and obfuscation.

DET plots illustrating detection performance in independence from ASV for both evasion and obfuscation are illustrated in Figure 4. The EER for evasion detection is 0% whereas that for obfuscation detection is 3%. ASV performance with combined obfuscation detection as a post-processing step [28] is illustrated for the IV-PLDA system in Figure 3. With the detector operating point set to the EER (Figure 4) there is almost no degradation in ASV performance (Figure 3) towards the low missed detection region. Corresponding EERs with integrated detection for all six systems are also illustrated in Table 1 and show EERs in the range of 3 to 4% in all cases.

5. CONCLUSIONS

This paper demonstrates the potential for surveillance and forensic speaker recognition systems to be manipulated. While ultimately they have the same effect, we introduce the notion of different, independent vulnerabilities to evasion and obfuscation which target either the detection or recognition modules. More importantly, the paper demonstrates the need and potential for new evasion and obfuscation detection countermeasures.

Our assessment shows large variations in system robustness with the most and least vulnerable GSL-FA and GMM-UBM systems showing EERs of 55% and 19% respectively when subjected to a trivial evasion attack. When subjected to obfuscation through voice conversion the most and least vulnerable GMM and IV-PLDA systems show EERs of 48% and 20% respectively. A new, generalised countermeasure shows that both evasion and obfuscation can be detected with reasonable accuracy; with EERs around 3 to 4% in all cases.

We acknowledge that the work presented in this paper is far from being exhaustive. Even if the trivial form of evasion examined in this paper may not overcome more sophisticated speech activity detection systems, and while the approach to obfuscation is perhaps beyond the means of the lay person, the observations reported here serve to highlight the need for further research to ensure that surveillance and forensic systems are adequately protected from both forms of subversion.

6. 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] N. Evans, T. Kinnunen, and J. Yamagishi, “Spoofing and countermeasures for automatic speaker verification,” in *Proc. Interspeech 2013*, Lyon, France, 2013.

[3] Q. Jin, A. R Toth, T. Schultz, and A. W. Black, “Voice convergin: Speaker de-identification by voice transformation,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2009, pp. 3909–3912.

[4] S. Yoon, J. Feng, and A.K. Jain, “Altered fingerprints: Analysis and detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 451–464, 2012.

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

[6] R Rodman, “Speaker recognition of disguised voices: a program for research,” in *Proceedings of the Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, COST250 Publishing Arm*. Citeseer, 1998, pp. 9–22.

[7] P. Perrot, G. Aversano, and G. Chollet, “Voice disguise and automatic detection: review and perspectives,” in *Progress in nonlinear speech processing*, pp. 101–117. Springer, 2007.

[8] H. J. Künzel, J. Gonzalez-Rodriguez, and J. Ortega-García, “Effect of voice disguise on the performance of a forensic automatic speaker recognition system,” in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2004.

[9] S. S. Kajarekar, H. Bratt, E. Shriberg, and R. de Leon, “A study of intentional voice modifications for evading automatic speaker recognition,” in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2006.

[10] C. Zhang and T. Tan, “Voice disguise and automatic speaker recognition,” *Forensic science international*, vol. 175, no. 2, pp. 118–122, 2008.

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

[12] P. Perrot, M. Morel, J. Razik, and G. Chollet, “Vocal forgery in forensic sciences,” in *Forensics in Telecommunications, Information and Multimedia*, pp. 179–185. Springer, 2009.

[13] H. Wu, Y. Wang, and J. Huang, “Blind detection of electronic disguised voice,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2013, pp. 3013–3017.

[14] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in *Proc. Interspeech*, 2007.

[15] W. M. Campbell, D. E. 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)*, may 2006, vol. 1, p. I.

[16] B. G. 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.

[17] 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.

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

[19] 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.

[20] 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.*, vol. 2004, pp. 430–451, Jan. 2004.

[21] 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*, 2008, vol. 5, p. 1.

[22] 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.

[23] F. Alegre, A. Amehraye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Washington DC, USA, 2013.

[24] 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.

[25] 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.

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

[27] Z. Wu, E.S. Chng, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” in *Proc. 13th Interspeech*, 2012.

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