EVASION AND OBFUSCATION IN AUTOMATIC SPEAKER VERIFICATION

Federico Alegre, Giovanni Soldi and Nicholas Evans

Multimedia Communications Department, EURECOM, Sophia Antipolis, France

{alegre,soldi,evans}@eurecom.fr

ABSTRACT

The potential for biometric systems to be manipulated through some form of subversion is well acknowledged. One such approach known as spoofing relates to the provocation of false accepts in authentication applications. Another approach referred to as obfuscation relates to the provocation of missed detections in surveillance applications. While the automatic speaker verification research community is now addressing spoofing and countermeasures, vulnerabilities to obfuscation remain largely unknown. This paper reports the first study. Our work with standard NIST datasets and protocols shows that the equal error rate of a standard GMM-UBM system is increased from 9% to 48% through obfuscation, whereas that of a state-of-the-art i-vector system increases from 3% to 20%. We also present a generalised approach to obfuscation detection which succeeds in detecting almost all attempts to evade detection.

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

1. INTRODUCTION

While biometrics systems play an increasingly ubiquitous role in person identification and security, the potential for the technology to be overcome through subversion is now well-acknowledged [1]. Subversion can take one of two general forms depending on the application: authentication or surveillance.

Spoofing relates to the authentication scenario and refers to the potential for an impostor to be accepted as an enrolled client. While the literature shows that the threat can be significant, countermeasures designed to detect spoofing attacks can also be highly effective [2]. While studied broadly in the case of other biometric modalities, research in automatic speaker verification (ASV) is only just beginning to gather pace [3].

This paper relates to the second, less-studied form of vulnerability involving surveillance applications. The problem here, referred to obfuscation, relates to the potential for a surveillance target to evade detection. While there is significant work in the literature which shows the vulnerability of other biometrics systems to obfuscation, e.g. [4], that in the case of ASV is largely unknown. Accordingly, we have set out to gauge ASV vulnerabilities to obfuscation and to investigate new detection approaches.

2. OBFUSCATION VERSUS SPOOFING

This section describes the difference between spoofing and obfuscation and relevant, past research.

2.1. Spoofing

Authentication applications involve identification or verification scenarios in which an enrolled client typically seeks the confirmation of their identity in order to gain access to protected resources. The likely attack in this scenario involves spoofing, which entails the impersonation or manipulation of an impostor's speech in order that it resembles that of an enrolled, target identity. The attack is thus intended to provoke a false acceptance, otherwise referred to as a Type II error.

The consideration of spoofing within the ASV community is relatively recent [3]. Spoofing attacks considered to date include impersonation [5], replay [6], speech synthesis [7], voice conversion [8] and artificial, non-speech signals [9]. Reports of false acceptance rates of well over 50% are not uncommon. Numerous spoofing countermeasures have thus emerged over recent years [10] and have potential to thwart most spoofing attacks.

2.2. Obfuscation

Surveillance applications involve the detection of one or more known speakers in a given audio recording, for example the detection of criminals in an intercepted telephone conversation. In such cases, persons of interest might disguise or manipulate their speech in order to evade detection [11, 12]. The intent here is to provoke a missed detection, otherwise referred to as a Type I error.

There is very little work in the literature relating to obfuscation, despite convincing arguments supporting the po-

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).

tential for obfuscation to overcome reliable recognition. The first relates to the notion of cooperation. In authentication, both naïve (zero-effort) impostor accesses and spoofing attacks involve cooperative interaction with the biometric system in as much that valid biometric samples are assumed to be collected. In contrast, the covert, surreptitious nature of surveillance often involves non-cooperative scenarios. The lack of cooperation can then impede the collection of biometric signals, i.e. through the use of discreet, far-field recording devices or low signal-to-noise ratios. The resulting lack of high-quality speech samples generally leads to higher missed detection rates.

The second relates to the notion of accessibility and effort. While spoofing requires a certain level of skill to faithfully imitate the speech of a *specific*, other individual, obfuscation requires only the hiding or disguise of the persons own, natural voice. Alternatively put, obfuscation involves the imitation of *any* other person's speech – a comparatively easier task.

2.3. Previous work

The work in [11, 13–15] 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 [12, 16] are also shown to overcome identification and verification systems. Again, however, this work uses non-standard, small datasets. The first work to detect disguised voice is reported in [17]. 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 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 obfuscation detection and analyse its impact on ASV performance.

3. EVALUATION

This section presents our work to assess the vulnerabilities of automatic speaker verification (ASV) to obfuscation. This section describes the different ASV systems, datasets and protocols used in this work, the particular approach to obfuscation, and experimental results.

3.1. ASV systems and feature extraction

We assessed the impact of obfuscation on six different ASV systems: (i) a standard Gaussian mixture model (GMM) with a universal background model (UBM); (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 [18]. 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 [19] whereas the third uses FA supervectors (GSL-FA) [20]. The i-vector system [21] employs intersession compensation with probabilistic linear discriminant analysis (PLDA) [22] with length normalisation [23]. From here on in it is referred to as IV-PLDA.

All ASV systems are based on the LIA-SpkDet toolkit [24] and the ALIZE library [25] and are directly derived from the work in [20]. They furthermore use a common UBM with 1024 Gaussian components, a common speech activity detector and feature parametrisation: linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy. Full details of all systems can be traced through [26].

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 obfuscation respectively.

To assess the potential impact of obfuscation, true-client tests are replaced with alternative speech data which aims to obfuscate reliable recognition. Any number of different approaches may be used to perform obfuscation. On account of its efficacy in spoofing [8] this paper considers voice conversion. The only difference between its application to study obfuscation instead of spoofing involves its application to client trials (instead of impostor trials) to provoke missed detections (instead of false accepts).

3.3. Voice conversion

Voice conversion is applied according to the Gaussian dependent filtering (GDF) approach proposed in [8]. It was originally used to assess ASV vulnerabilities to spoofing by transforming impostor test utterances towards the speech of target speakers. 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

System	EER (%)		minDCF ×100	
	Baseline	Obfus.	Baseline	Obfus.
GMM-UBM	8.7	34.2	4.14	10.06
GSL	8.0	19.6	3.52	9.00
GSL-NAP	6.8	18.9	2.65	7.70
FA	5.6	28.7	2.32	9.39
GSL-FA	6.4	16.6	2.48	7.19
IV-PLDA	3.0	8.0	1.15	4.03

Table 1. ASV performance with and without obfuscation through voice conversion towards the UBM. Results shown in terms of EER and minDCF $\times 100$.

	EER (%)		minDCF ×100	
	GMM	IV-PLDA	GMM	IV-PLDA
Baseline	8.7	3.0	4.14	1.15
UBM	34.2	8.0	10.06	4.03
Random	34.5	12.0	10.02	7.09
Dissimilar	47.7	20.0	10.15	9.89

Table 2. GMM and IV-PLDA performance with different approaches to obfuscation with voice conversion towards the UBM, a random speaker and the most dissimilar speaker in the NIST dataset. Results shown in terms of EER and minDCF $\times 100$.

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:

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

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 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 [8,27,28].

3.4. Results

We investigated three different approaches involving the conversion of all target tests towards: (i) the universal background model; (ii) a randomly selected speaker and (iii) the most dissimilar speaker in the NIST dataset. Conversion towards the UBM aims to increase the likelihood of the inverse hypothesis and thus to decrease the likelihood ratio. Conversion towards a randomly selected speaker is intended to reflect the averaged effect of conversion whereas conversion towards the most dissimilar speaker (that for which the resulting likelihood score is the lowest) is intended to reflect a worst-case scenario.

Table 1 illustrates the effect of obfuscation with conversion towards the UBM. Results are presented in terms of the equal error rate (EER) and the minimum decision cost function (minDCF). We discuss only the former in the following. The GMM-UBM system is the most vulnerable and shows a degradation from 9% EER to 34% EER. The FA and three GSL-based systems show moderate vulnerability whereas the IV-PLDA system is the most robust; the EER increases from 3% to only 8%.

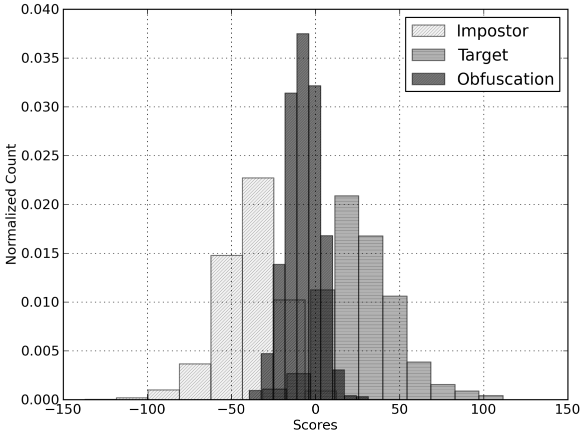

Fig. 1. IV-PLDA score distributions for impostor (left-most), target (right-most) and obfuscation trials with conversion towards a random speaker.

Table 2 illustrates a comparison of performance for the three different approaches to voice conversion and for the least and most robust systems. With conversion towards a randomly selected speaker the EERs of the GMM and IV-PLDA systems increase to 35% and 12% respectively. When conversion is performed towards the most dissimilar speaker, then EERs increase further to 48% and 20%.

Figure 1 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 towards a random speaker is effective in decreasing the resulting likelihood scores for target tests; the degree of overlap with the impostor distribution is higher than it is for the target distribution. Detection error

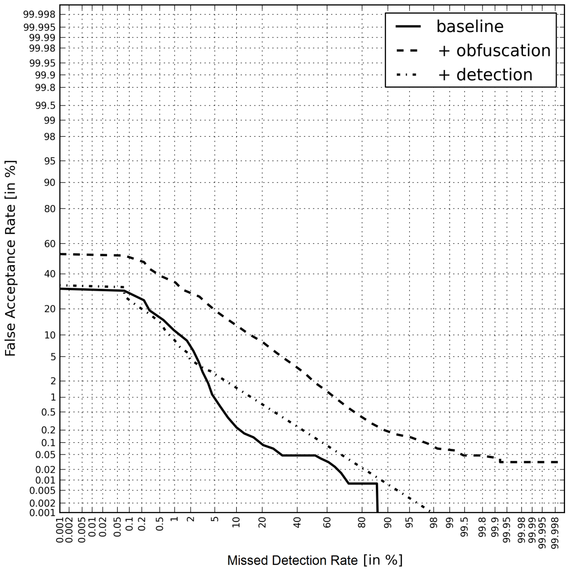

Fig. 2. DET profiles illustrating IV-PLDA performance for the baseline, obfuscation trials with conversion towards a random speaker and performance with obfuscation detection.

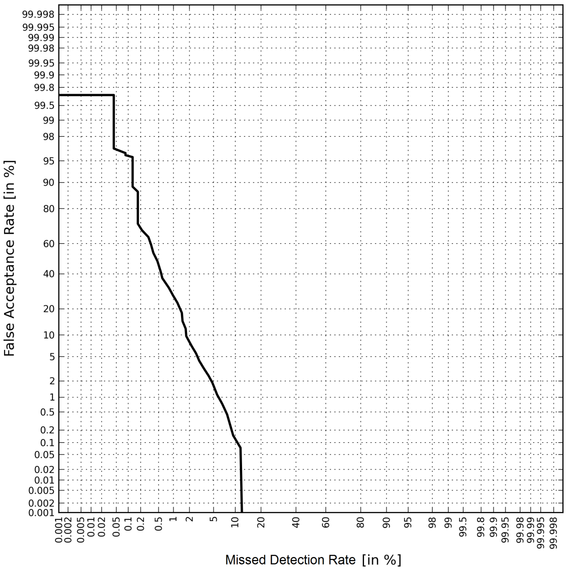

Fig. 3. A DET profile illustrating detection performance independently from ASV.

trade-off (DET) profiles[1] for the IV-PLDA systems are illustrated in Figure 2. Profiles for the baseline and obfuscation (conversion to random speaker) show that the system is vulnerable across the full range of operating points but slightly more robust in the area of low missed detections.

4. DETECTION

Various different approaches to detect converted voice have been reported in the literature. All involve the study of spoofing and the detection of processing artifacts indicative of manipulated, converted speech, e.g. the absence of natural-speech phase [29] and reduced short-term dynamic variability [30].

These approaches are, however, dependent on the specific approach to voice conversion and thus have limited practical application. The work in [26] investigated a more generalised solution with the potential to detect previously unseen approaches to voice conversion, or indeed other approaches to spoofing generally. 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.

A DET plot illustrating detection performance in independence to ASV is illustrated in Figure 3 and shows an EER of 3%. ASV performance with combined obfuscation detection as a post-processing step [30] is illustrated in Figure 2. With the operating point set to the EER (Figure 3) there is almost no degradation in ASV performance on account of obfuscation (Figure 2) towards the low missed detection region.

5. CONCLUSIONS

This paper assesses the potential for surveillance targets to evade detection through automatic speaker verification and evaluates the potential for obfuscation detection.

Our assessment shows variations in system robustness with the GMM-UBM and IV-PLDA systems showing EERs of up to 48% and 20% when subjected to obfuscation. While this work shows the tangible potential for surveillance targets to evade detection, we fully acknowledge that the work suffers from some of the same shortcomings as all previous work in spoofing. They include the consideration of high-technology voice conversion which is most likely beyond the grasp of the average surveillance target, and also the consideration of only one approach to obfuscation.

While the work presented in this paper thus arguably over-exaggerates vulnerabilities to obfuscation, we nonetheless show how a generalised form of detection succeeds in identifying almost all obfuscation tests.

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

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] "EU FP7 Project – 'TABULA RASA' – www.tabularasa-euproject.org," 2010–2014.

[3] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech 2013*, Lyon, France, 2013.

[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] 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.

[6] 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*, 1999, vol. 3, pp. 1211–1214.

[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] 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.

[9] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Proc. 12th EUSIPCO*, 2012.

[10] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*, S. Marcel, S. Z. Li, and M. Nixon, Eds. Springer, 2014.

[11] 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.

[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. 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.