A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures using Local Binary Patterns

Federico Alegre, Asmaa Amehraye and Nicholas Evans
Multimedia Communications Department, EURECOM
Sophia Antipolis, France
{alegre,fillatre,evans}@eurecom.fr

Abstract

The vulnerability of automatic speaker verification systems to spoofing is now well accepted. While recent work has shown the potential to develop countermeasures capable of detecting spoofed speech signals, existing solutions typically function well only for specific attacks on which they are optimised. Since the exact nature of spoofing attacks can never be known in practice, there is thus a need for generalised countermeasures which can detect previously unseen spoofing attacks. This paper presents a novel countermeasure based on the analysis of speech signals using local binary patterns followed by a one-class classification approach. The new countermeasure captures differences in the spectro-temporal texture of genuine and spoofed speech, but relies only on a model of the former. We report experiments with three different approaches to spoofing and with a state-of-the-art i-vector speaker verification system which uses probabilistic linear discriminant analysis for intersession compensation. While a support vector machine classifier is tuned with examples of converted voice, it delivers reliable detection of spoofing attacks using synthesized speech and artificial signals, attacks for which it is not optimised.

1. Introduction

In the context of biometric authentication, spoofing refers to the presentation of a falsified or manipulated trait to the sensor of a biometric system in order to provoke a high score and illegitimate acceptance. Unless the biometric system is equipped with appropriate spoofing countermeasures, this threat is common to all biometric modalities. For example, face recognition systems can be spoofed with a photograph [1], whereas fingerprint recognition systems can be spoofed with a fake, gummy finger [2].

Automatic speaker verification (ASV) systems are also vulnerable to spoofing attacks with varying degrees of sophistication. Impersonation [3, 4], replayed speech [5, 6], synthesised speech [7, 8], voice conversion [9–12] and artificial signals [13] have all been shown to provoke significant increases in the false acceptance rate of state-of-the-art ASV systems. More often than not, authentication performance is then comparable to that expected by chance.

Recently the research community has started to investigate spoofing actively. Although there is one notable exception in face recognition [14], due to the novelty of such work there are currently no standard large-scale datasets, protocols or metrics for the evaluation of spoofing countermeasures. In consequence, it is still common practice for countermeasures to be developed using closed, purpose-collected datasets and, more critically, often with full prior knowledge of the spoofing attack. This assumption is unrealistic; in practice the spoofing attack can never be known and then the performance of existing countermeasures in practical scenarios is unknown.

Recent work shows the potential impact of using prior knowledge. For instance, de Freitas Pereira et al. [15] showed that state-of-the-art spoofing countermeasures for face recognition do not generalise well to forms of spoofing not considered during development. Similar behaviour can be expected in speaker recognition. Countermeasures based on phase [16–18] and prosodic features [19, 20] can be used very successfully to detect voice conversion and speech synthesis attacks. It is likely, however, that they will be overcome by the particular approach to voice conversion investigated in [11] which modifies only the spectral slope of a speech utterance while retaining the original phase and pitch of the original, genuine speech signal. Spoofing thus remains very much an open problem.

This paper reports our work to develop a generalised countermeasure for speaker recognition. New contributions are two-fold. First, we introduce a new feature for spoofing detection based on the analysis of conventional speech parameterisations using Local Binary Patterns (LBPs) [21] and second, we present the first one-class classification approach to ASV spoofing detection. Experiments using

a state-of-the-art i-vector / probabilistic linear discriminant analysis (PLDA) ASV system and three different approaches to spoofing show that the new system is less dependent on prior knowledge; while the countermeasure is optimised for the detection of converted voice, it is also effective in detecting synthesized speech and artificial signals.

The remainder of this paper is organized as follows. Spoofing attacks and the new countermeasure are presented in Sections 2 and 3, respectively. Experimental work is described in Section 4. Our conclusions are presented in Section 5.

2. Spoofing attacks

In this section we describe our approach to voice conversion, speech synthesis and attacks with artificial signals.

2.1. Voice Conversion

All work involving voice conversion was performed with our own implementation of the approach originally proposed in [11]. It was developed to test the limits of ASV when the vocal tract information in the speech signal of a spoofer is converted towards that of another, target person. At the frame level, the speech signal of a spoofer denoted by $y(t)$ is filtered in the spectral domain as follows:

$$Y'(f) = \frac{|H_x(f)|}{|H_y(f)|} Y(f) \tag{1}$$

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 speaker in a spectral-slope sense. As we show later, this 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 utilized 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:

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

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 synthesized from converted frames with a standard overlap-add technique. Full details can be found in [11, 22, 23].

2.2. Speech synthesis

Spoofing attacks with speech synthesis were implemented using the hidden Markov model (HMM)-based Speech Synthesis System (HTS)[1] and the specific approach described in [24]. Parameterisations include 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) along with their delta and acceleration coefficients. Acoustic spectral characteristics and duration probabilities are modeled using multispace distribution hidden semi-Markov models (MSD-HSMM) [25]. 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) [26]. Finally, time domain signals are synthesized 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 [27].

2.3. Artificial signals

Artificial signal attacks are based on the algorithm reported in [13]. It is based on a modification of the voice conversion algorithm presented in Section 2.1.

Let $S = \{c_1, ..., c_n\}$ be a short sequence of consecutive speech frames selected from an utterance of the targeted speaker. The algorithm seeks a new sequence of speech frames S^* which maximises the score of a given ASV system and thus the potential for spoofing.

Each frame $c(t)$ belonging to S is initially transformed in the frequency domain with voice conversion where we now have:

$$C'(f) = \frac{|H^*_c(f)|}{|H_c(f)|} C(f) \tag{3}$$

Optimisation is then applied to identify a set of filters $H^*_S = \{H^*_{c_1}(f), H^*_{c_2}(f), ..., H^*_{c_n}(f)\}$. Instead of estimating each filter independently using Equation 2, however, the set of filters is jointly optimized using a genetic algorithm. Full details are presented in [13].

3. LBP spoofing countermeasure

In common with all previous work in spoofing countermeasures for ASV, that reported in this paper is conducted with full prior knowledge of a single, specific spoofing attack, namely voice conversion. The paper also reports the first work in speaker verification to assess the generality

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

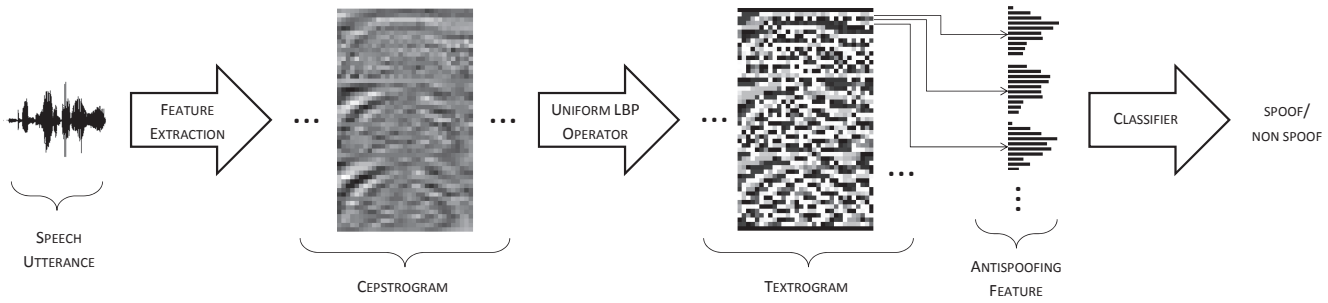


Figure 1. Application of uniform LBP analysis to a cepstrogram to obtain the so-called textrogram. Non-uniform patterns are discarded and normalised histograms of the remaining uniform LBPs in each row are concatenated to form the new anti-spoofing feature.

of countermeasures to unseen spoofing attacks. Here we also assess countermeasure performance in detecting attacks with synthesized speech and artificial signals. No knowledge of these algorithms was used intentionally during development.

3.1. LBP features

Based on our previous work [28], we hypothetise that genuine speech can be distinguished from spoofed speech according to differences in the spectro-temporal 'texture'. The motivation stems from the assumption that, while lower-level spectral representations can be synthesized with relative ease, higher-level, longer-term spectro-termporal information is considerably more difficult to spoof. The new countermeasure reported in this paper is based on the application of a standard approach to texture analysis known as Local Binary Patterns [21].

As illustrated in Figure 1, LBP analysis is applied to a 2-dimensional 'image' of a speech utterance, where here the image is a cepstrogram formed from the concatenation of traditional cepstral features, including standard velocity and acceleration features. LBPs are determined for each pixel in the cepstrogram using any appropriate LBP operator, thus resulting in a new matrix of reduced dynamic range, here referred to as a 'textrogram'. The dimensions of the textrogram are determined by the number of components in each feature vector and the duration of the speech signal. The textrogram captures short-time, spectro-temporal feature motion beyond that in conventional dynamic parametrizations.

Classification of speech utterances as either genuine speech or spoofed speech is based upon a new set of features extracted from the textrogram. A histrogram of LBPs is constructed for each row of the textrogram. The set of histograms are independently normalised and the new anti-spoofing feature is formed from the concatenation of each histogram into a single super-vector. Example cepstrograms (left) and textograms (right) are illustrated in Figureřeffig:LBPexamples for both genuine speech (top) and a spoofed attack through voice voice conversion (bottom). While a certain level of smoothing is detectable in the cepstrograms, differences in the textrograms are more pronounced (although not immediately obvious by eye) and point to the potential of the new approach to detect spoofing.

3.2. Classification and integration

The problem of generality in the context of multiple, unknown spoofing attacks has been investigated previously for face recognition. Chingovska et al. [29] showed the benefit of attack-optimised LBP-operators. Recent efforts to develop generalised countermeasures have accordingly investigated multi-class classifiers to deal with the variation in possible attacks. Solutions to anti-spoofing then entail the fusion of a number of binary classifiers where each classifier is optimised to detect a specific attack or spoofing indicator. The same technique may also be used to enhance robustness to a single, specific attack e.g. the combination of motion and texture analysis for face anti-spoofing [30].

Binary spoofing detectors are generally independent of the biometric system and typically trained using both genuine data (negative samples) and examples of spoofed data (positive samples). The main drawback of such an approach is the over fitting to spoofing attacks seen in the training data and thus the lack of generalization to attacks previously unseen [15]. While additional classifiers can be trained and integrated when new attacks are identified, clearly this leads to increased system complexity.

In comparison to some other biometric modalities, spoofing and countermeasure research in ASV is far less advanced. While attacks from impersonation, replay, speech synthesis and voice conversion are all known, there is a high degree of variation in specific algorithms and there are certainly other forms of attack yet to be identified. Current work in spoofing countermeasures for ASV thus optimistically biases results to known attacks and specific algorithms. While the true extent of spoofing in the context of ASV is yet to be fully understood, and in any case, there is

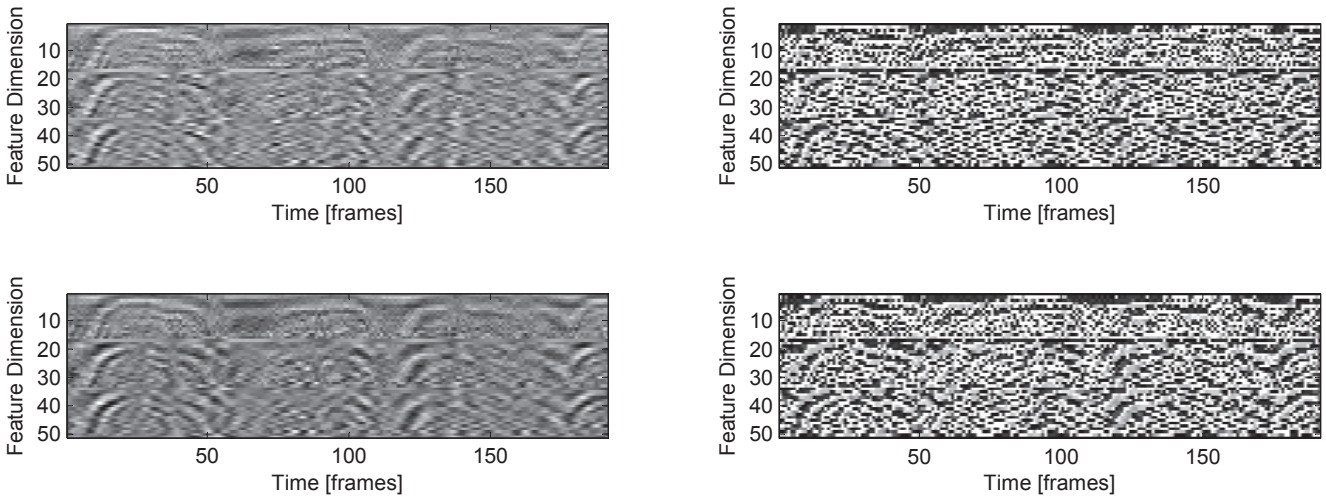


Figure 2. On the left: Example of concatenated feature vectors extracted from 193 consecutive speech frames (approximately 2 seconds of continue speech) of real speech (above) and its converted version (below). On the right: uniform LBP operator applied to feature vectors. Note that each 'image' is comes from aprox. 2.5 min of speech (around 10000 frames)

thus a need for generalised countermesures.

Accordingly we have recently pursued a one-class classification approach to detect spoofing. One-class classifiers differ from two-class and multi-class classifiers in that only data from one class is used for training, and therefore classifiers are designed to distinguish between the one known class and any other which is unseen during training. Applications of these classifiers are related to anomaly/outlier detection. One class support vector machine (SVM) classifiers are usually preferred, where the idea is to minimize the volume of the hypersphere which contains the training data.

In Section 4 we report experiments with this approach to classification, among others. In all cases the proposed countermeasure is integrated with the ASV system as an independent post processing step, in equivalent fashion to the work in [8, 17, 28].

4. Experimental work

Here we report experiments to assess the performance of the new countermeasure using a state-of-the-art ASV system and public, standard datasets.

4.1. ASV systems

In previous work [28] we investigated the vulnerability to spoofing of five different ASV systems, from a standard GMM-UBM system to a GMM supervector linear kernel (GSL) system, also combined with channel compensation techniques such as nuisance attribute projection (NAP) and Factor Analysis (FA). New to this contribution is the consideration of an i-vector system, the current state of the art in speaker verification [31]. The i-vector system uses FA to model session and speaker variability at the front-end by means of a so-called total variability matrix. The setup involves mixtures of 1024 Gaussian components and i-vectors with 400 dimensions. The total variability matrix estimation and i-vector extraction is performed using the ALIZE toolkit [32] and the LIA-RAL framework [33]. Unwanted variability is handled through Probabilistic Linear Discriminant Analysis (PLDA) compensation [34] with length normalization [35].

4.2. Protocols and metrics

The male subset of the 2005 NIST Speaker Recognition Evaluation (NIST'05) dataset is used for development and the NIST'06 dataset is used for evaluation. As in [28], all experiments relate to the 8conv4w-1conv4w condition – where one conversation provides an average of 2.5 minutes of speech (one side of a 5 minute conversation). To ensure no overlap between data used for ASV or countermeasure development and data used for voice conversion, only one of the 8 training conversations is ever used for the former whereas the other 7 are set aside for learning voice conversion models.

Due to the significant amount of data necessary to estimate the total variability matrix T, the NIST'06 dataset was used as background data during development and the NIST'05 dataset was used during evaluation. In both cases the background datasets were augmented with the NIST'04 and NIST'08 datasets. In both cases, matrices are estimated with approximately 11,000 utterances from 900 speakers, while independence between development and evaluation experiments is always respected. Finally, speech data used for UBM learning comes either from the NIST'04 or NIST'08 datasets depending on whether the resulting GMM is used for ASV or for spoofing respectively.

As illustrated in Figure 1, standard NIST protocols dic-

Dataset	NIST'05 (dev)	NIST'06 (eval)
Speakers	201	298
Client tests	984	1344
Impostor tests	9862	12648

Table 1. Number of target and impostor trials in the development and evaluation datasets.

tate in the order of 1,000 true client tests and 10,000 impostor tests for development and evaluation datasets. In all spoofing experiments, both the number of true client tests and impostor tests are the same as for the baseline, but the speech samples of each impostor test are converted toward the corresponding client model. 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. The countermeasure is assessed independently of ASV, also in terms of EER. Also reported are the false acceptance rate (FAR) for a fixed false rejection rate (FRR) of 10%.

4.3. Spoofing attacks

The voice conversion system is identical to that described in [28], while for artificial signal generation we adopted the setup reported in [36]. Speech synthesis attacks were implemented using the voice cloning toolkit[2] with a default configuration. We used standard speaker-independent models provided with the toolkit which were trained on the EMIME corpus [37]. Synthesized speech is generated using the transcripts of the original impostor utterances.

While it is admittedly not representative of real scenarios, we assess countermeasure performance in a worst case scenario, where the attacker/spoofer has full prior knowledge of the ASV system. Voice conversion and artificial signal attacks thus use the same features used for ASV. We do not consider this to be a significant weakness since we note that other work has observed only minor differences in vulnerability when the ASV systems used to effect spoofing are different [23].

4.4. LBP countermeasure

LBP analysis is applied to cepstrograms composed of 51 coefficients: 16 LFCCs and energy plus their corresponding delta and delta-delta coefficients. Frame blocking is the same as for ASV systems (although different frame lengths do provide similar results). We take into account only those frames determined to contain speech, i.e. those also used for ASV.

We performed experiments with $LBP_{4,1}$, $LBP_{8,1}$, $LBP_{8,2}$, and $LBP_{16,2}$ operators and their uniform versions using the LBP Matlab implementation made available by The University of Oulu[3]. Our best results were obtained with a $LBP^{u2}_{8,1}$ operator considering only the 58 possible uniform patterns. Histograms of LBPs are created for all but the first and last frames, thereby obtaining a $58 \times (51 - 2) = 2842$ length feature vector.

We assess three different classifiers. In all cases attacks with speech synthesis and artificial signals represent the universe of unknown attacks. For the first two, one-class approaches, only converted voice was used for optimisation (tuning as opposed to training). The first classifier is one-class[4], speaker-dependent approach whereby scores correspond to the comparison of the LBP feature vector extracted from the input utterance to that of the target client (from the ASV training dataset) using a histogram intersection kernel. The second classifier is a one-class, speaker-independent SVM approach where scores correspond to the comparison of the input utterance to the set of LBP features extracted from all utterances in the NIST'04 and NIST'08 datasets (approximately 8000 utterances). The third classifier is a two-class SVM where each of the two models are trained on the same genuine speech as the second classifier and the 9892 converted voice utterances in the development set respectively. All SVM classifiers are implemented using the LIBSVM[5] library [38] and are tuned using only geunine speech or converted voices in the development set.

4.5. Results

Here we report the performance of the baseline system, the effect on performance of spoofing attacks with voice conversion, speech synthesis and artificial signals, and then improvements in robustness afforded through the new, LBP-based countermeasure. Only results obtained on the evaluation dataset are reported. Detection error trade-off (DET) profiles for each of the three attacks are presented in Figure 4. In each of the three plots, the effect of spoofing is illustrated by the differences between the 1^{st} (baseline) and 3^{rd} (baseline under attack) profiles.

Results are summarised in Table 2 in terms of EER (%, 1^{st} column) and also in terms of FAR (%, 2^{nd} column) for a fixed FRR of 10%. The baseline system produces a competitive performance of 3% EER. This corresponds to an FAR of 0.2%. While the i-vector system is more robust than alternative systems considered in our previous work [28], all attacks nonetheless provoke significant increases in both EER and FAR. The most significant increase is observed for attacks with voice conversion where the EER increases to 22% and the FAR increases to 55%. We note that speech

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

[3] http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

[4] Only real speech is used for modeling.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Dataset	ASV EER	ASV FAR	ASV + CM FAR
Baseline	3	0.2	–
Voice conversion	22	55	4.1
Speech synthesis	10.4	11	0
Artificial Signals	7.6	4.8	0

Table 2. ASV performance in terms of EER (%) and FAR (%) for a fixed FRR of 10% for the baseline system and when subjected to spoofing.

synthesis and artificial signal attacks are less effective than voice conversion since they target ASV systems at the feature level rather than at the GMM distribution or component level.

Results for the new LBP-based countermeasure and each of the three different classifiers are illustrated in Table 3. For the one-class SVM classifier, we obtained our best results with a radial kernel basis function, while a linear kernel gave better results for the two-class classifier. As expected, compared to the one-class classifiers, the two-class classifier offers the best performance for the condition on which it is optimised (voice conversion). Here the EER is 0%. However, for the two spoofing attacks not seen during optimisation, performance is poor. Since the binary SVM classifier is not designed to manage 'outliers' it is perhaps not surprising in this case that EERs increase rather than decrease. While the one-class classifiers do not perform as well as the two-class classifier for voice conversion spoofing attacks, EERs of 8% and 5% are only marginally higher than the baseline EER of 3%. More importantly, the one-class classifiers are seen to generalise well to synthesised speech and artificial signals. Here the EERs are all less than 1%.

We see from Table 3 that the best overall performance is obtained with the one-class SVM classifier. A detection error trade-off (DET) profile which illustrates the performance of the one-class SVM classifier independently from ASV is illustrated[6] in Figure 3. Here we see EERs of 5%, 0.1% and 0% for voice conversion, speech synthesis and artificial signals respectively. We note that approximately 0.2% of true tests give very bad scores (flattening effect towards the top left of the profile for voice conversion in Figure 3), which is again expected for such a classifier.

Finally, we consider overall performance when the spoofing countermeasure is integrated with the ASV system. The effect of the new countermeasure on licit transactions is illustrated in each plot of Figure 4 by the difference between the 1^{st} (baseline) and 2^{nd} (baseline + countermeasure) profiles. The effect on spoofed transactions is

[6]TABULA RASA scoretoolkit: http://publications.idiap.ch/downloads/reports/2012/Anjos_Idiap-Com-02-2012.pdf

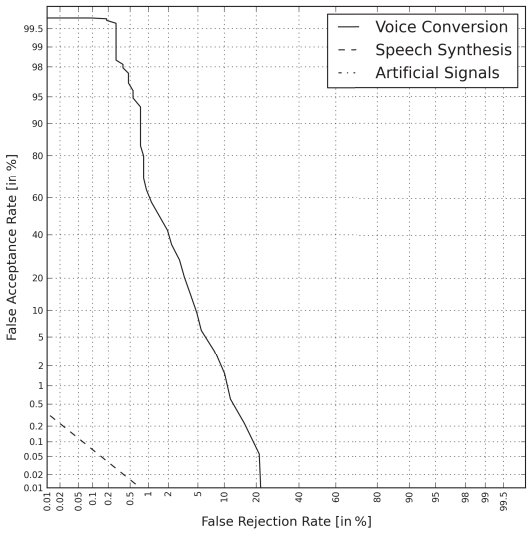


Figure 3. DET profiles illustrating countermeasure performance independently from ASV. The profile for artificial signals in not visible since the EER is 0%.

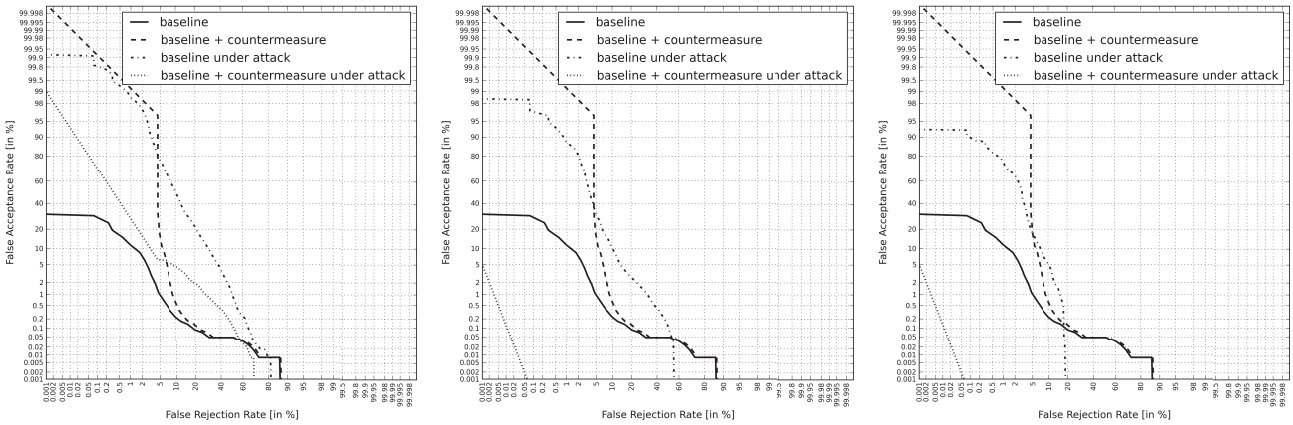
Classifier Attack	1-class spk-dep	1-class SVM	2-class SVM
Voice Conversion	8	5	0
Speech Synthesis	1	0.1	56
Artificial Signals	0	0	25

Table 3. Countermeasure performance in terms of EER (%) for the three different classifiers and three different spoofing attacks.

illustrated by the difference between the 3^{rd} (baseline under attack) and 4^{th} (baseline + countermeasure under attack) profiles. Results are summarised in Table 2 which shows the protection offered through the one-class SVM countermeasure when its operating point is set to its ERR and it is used to identify spoofed speech prior to ASV. We observe that FARs drop from 55%, 11% and 4.8% to 4.1% for voice conversion and 0% for speech synthesis and artificial signals respectively. In all cases the FARs are then much more comparable to that of the baseline and the new countermeasure thus generalises well to unseen spoofing attacks.

5. Conclusions and future work

In practice, one can expect biometrics systems to be subjected to a wide variety of unpredictable spoofing attacks. Accordingly, there is a need for generalised spoofing countermeasures with the potential to detect attacks for which they have not been optimised. This paper addresses this issue to some extent. The new countermeasure based on local binary patterns is shown to give the best reported results to date for converted voice spoofing attacks. We also show the potential of a one-class classification approach which allows for the new countermeasure to be applied successfully



(a) Voice Conversion (b) Speech Synthesis (c) Artificial Signals

Figure 4. DET profiles for the baseline i-vector system with and without the proposed countermeasure. The three figures represent system performance with and without spoofing attacks with voice conversion, speech synthesis and artificial signals. In all cases the countermeasure operating points is set to the ERR (5%).

to the detection of alternative spoofing attacks not seen during development. While the paper thus demonstrates the potential for generalised countermeasures, we have worked with such signals for some time. Accordingly we have nonetheless benefited to a certain extent from our familiarity with the so-called 'unseen' attacks . Formal evaluations with standard corpora, protocols and metrics are therefore needed to stimulate the research of spoofing countermeasures under properly controlled settings reflective of practical use case scenarios and with genuinely unseen and varying attacks. The development of effective countermeasures will then be extremely challenging.

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) and the ALIAS project (AAL-2009-2-049 - co-funded by the EC, the French ANR and the German BMBF).

References

[1] N. M. Duc and B. Q. Minh, "Your face is not your password face authentication bypassing lenovo–asus–toshiba," *Black Hat Briefings*, 2009.

[2] J. Galbally-Herrero, J. Fierrez-Aguilar, J. D. Rodriguez-Gonzalez, F. Alonso-Fernandez, J. Ortega-Garcia, and M. Tapiador, "On the vulnerability of fingerprint verification systems to fake fingerprints attacks," in *Proceedings 40th Annual IEEE International Carnahan Conferences Security Technology*. IEEE, 2006, pp. 130–136.

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

[4] M. Farrús, M. Wagner, J. Anguita, and J. Hern, "How vulnerable are prosodic features to professional imitators?," in *Proc. Odyssey IEEE Workshop*, 2008.

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

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

[7] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999.

[8] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey IEEE Workshop*, 2010.

[9] B.L. Pellom and J.H.L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. ICASSP*, 1999, vol. 2, pp. 837–840.

[10] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP : Indexation in a Client Memory," in *Proc. ICASSP*, 2005, vol. 1, pp. 17 – 20.

[11] 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.

[12] T. Kinnunen, 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 *Proc. ICASSP*, 2012, pp. 4401–4404.

[13] 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.

[14] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *International Joint Conference on Biometrics (IJCB),*. IEEE, 2011, pp. 1–7.

[15] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?," *6th IAPR International Conference on Biometrics (ICB)*, 2013.

[16] 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.

[17] Z. Wu, T. Kinnunen, E.S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–5.

[18] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. ICASSP*, 2011, pp. 4844–4847.

[19] A. Ogihara and A. Shiozaki, "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, 2005.

[20] P.L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. 13th Interspeech*, 2012.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[22] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. Odyssey IEEE Workshop*, 2006, pp. 1–6.

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

[24] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker adaptive HMM based Text-to-Speech Synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[25] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," in *Proc. ICASSP*, 1985, pp. 5–8.

[26] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," *IEEE transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[27] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[28] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *To appear in Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[29] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*. IEEE, 2012, pp. 1–7.

[30] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," *6th IAPR International Conference on Biometrics (ICB)*, 2013.

[31] 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.

[32] 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. Odyssey IEEE Workshop*, 2008, vol. 5, p. 1.

[33] 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.

[34] Peng Li, Yun Fu, Umar Mohammed, James H Elder, and Simon JD Prince, "Probabilistic models for inference about identity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 144–157, 2012.

[35] 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.

[36] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals," in *Proc. 13th Interspeech*, 2012.

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

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.