ON THE VULNERABILITY OF AUTOMATIC SPEAKER RECOGNITION TO SPOOFING ATTACKS WITH ARTIFICIAL SIGNALS

Federico Alegre[1], Ravichander Vipperla[1], Nicholas Evans[1] and Benoît Fauve[2]

[1]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
[2]ValidSoft Ltd, London, UK
{*alegre,vipperla,evans*}*@eurecom.fr, benoit.fauve@validsoft.com*

ABSTRACT

Automatic speaker verification (ASV) systems are increasingly being used for biometric authentication even if their vulnerability to imposture or spoofing is now widely acknowledged. Recent work has proposed different spoofing approaches which can be used to test vulnerabilities. This paper introduces a new approach based on artificial, tone-like signals which provoke higher ASV scores than genuine client tests. Experimental results show degradations in the equal error rate from 8.5% to 77.3% and from 4.8% to 64.3% for standard Gaussian mixture model and factor analysis based ASV systems respectively. These findings demonstrate the importance of efforts to develop dedicated countermeasures, some of them trivial, to protect ASV systems from spoofing.

Index Terms— biometrics, speaker verification, spoofing, imposture

1. INTRODUCTION

The state-of-the-art in automatic speaker verification (ASV) has advanced rapidly in recent years. Surprisingly, however, there has been relatively little work in the development of countermeasures to protect ASV systems from the acknowledged threat of spoofing. Otherwise referred to as the direct, sensor-level or imposture attacks of biometric systems [1], spoofing is generally performed by the falsification or impersonation of a biometric trait and its presentation to the biometric system. Since their natural appeal lies in automated, unattended scenarios, speaker recognition systems are particularly vulnerable to spoofing attacks.

Examples of ASV spoofing include impersonation [2, 3], replay attacks [4, 5], voice conversion [6–8] and speech synthesis [9, 10]. All of these approaches can be used to bias the distribution of impostor scores toward the true client or target distribution and thus to provoke significant increases in the false acceptance rate of ASV systems.

Through EURECOM's participation in the European Tabula Rasa project[1] we have investigated the vulnerability of ASV systems to a previously unconsidered spoofing attack in the form of artificial, tone-like signals. Given some speaker-specific training data, artificial signals can be synthesized and injected into an attackers natural voice signal or, as is the case investigated in this paper, used on their own to boost the ASV system score. With such signals having the potential to pass both energy-based and pitch-based voice activity detection systems, this new attack vector presents a serious threat if no specific countermeasures, some of them trivial, are introduced in the design of the ASV systems. The lack of any relevant prior work and our own experimental results lead us to believe that the threat from artificial signals is underestimated and warrants wider attention.

The paper is organized as follows. Section 2 presents the algorithm used to generate artificial signals to test vulnerabilities to spoofing attacks. Experimental work which aims to gauge the threat is described in Section 3. Finally our conclusions and ideas for future work are presented in Section 4.

2. ARTIFICIAL SIGNAL ATTACKS

Here we describe the approach to generate artificial spoofing signals. It is based on previous work in voice conversion by other authors [7], the relevant components of which are described first.

2.1. Voice conversion

Source-filter models are widely used in speech signal analysis and form the basis of a recent approach to voice conversion. According to the notation in [7] a speech signal Y of n frames

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), the ALIAS project (AAL-2009-2-049 - co-funded by the EC, the French ANR and the German BMBF) and by a Futur et Ruptures award from Institut TELECOM.

[1]The EU FP7 Tabula Rasa project (`www.tabularasa-euproject.org`) aims to develop new spoofing countermeasures for different biometric modalities including voice.

$\{y_1, ..., y_n\}$ is alternatively represented in the spectral domain according to the standard source-filter model:

$$Y(f) = H_y(f)S_y(f) \quad (1)$$

where $H_y(f)$ is the vocal tract transfer function of Y and $S_y(f)$ is the Fourier transform of the excitation source.

The separation of excitation and vocal tract information facilitates the conversion of an imposter's speech signal toward the speech of another, target speaker. In [7] a converted voice Y' is obtained by replacing an impostor transfer function $H_y(f)$ in (1) with that of a target speaker $H_x(f)$ according to:

$$Y'(f) = H_x(f)S_y(f) = \frac{H_x(f)}{H_y(f)}Y(f) \quad (2)$$

If the phase of the impostor signal is left unaltered, Y is thus mapped toward X in the spectral-slope sense by applying to $Y(f)$ a filter:

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \quad (3)$$

The transfer functions above are estimated according to:

$$H_y(f) = \frac{G_y}{A_y(f)}, \text{ and} \quad (4)$$

$$H_x(f) = \frac{G_x}{A_x(f)} \quad (5)$$

where $A_y(f)$ and $A_x(f)$ are the Fourier transforms of the corresponding prediction coefficients and G_y and G_x are the gains of the corresponding residual signals. While $H_y(f)$ is obtained directly from Y, $H_x(f)$ is estimated by using two parallel sets of Gaussian mixture models (GMM) of the target speaker. Full details are presented in [7].

2.2. Artificial Signals

Our approach to test the vulnerabilities of ASV systems to spoofing combines voice conversion and the notion of so-called replay attacks, where a genuine-client recording is replayed to a biometric sensor, here a microphone. Particularly if it is equipped with channel compensation routines, then it is entirely possible that an ASV system may be overcome through the replaying of a client speech signal X; this is a conventional replay attack. However, certain short intervals of contiguous frames in X, e.g. those corresponding to voiced regions, will give rise to higher scores or likelihoods than others. The probability of a replay attack overcoming and ASV system can thus be increased by selecting from X only those components or frames which provoke the highest scores. The resulting signal will not sound anything like intelligible speech but this is of no consequence if we assume, as is generally the case, that the ASV system in question uses only energy and/or pitch-based speech activity detection (SAD) and does not incorporate any form of speech quality assessment.

Here we consider an attack based upon the extraction and replaying of a short interval or sequence of frames in $X = \{x_1, ..., x_m\}$ which gives rise to the highest scores. Let $T = \{t_1, ..., t_n\}$ be such an interval short enough so that all frames in the interval provoke high scores, but long enough so that relevant dynamic information (e.g. delta and acceleration coefficients) can be captured and/or modelled. In order to produce a replay recording of significant duration, T can be replicated and concatenated any number of times to produce an audio signal of arbitrary length. In practice the resulting concatenated signal is an artificial, or tone-like signal which reflects the pitch structure in voiced speech.

Even though such signals can be used themselves to test the vulnerabilities of ASV systems, their limits can be more thoroughly tested by enhancing the above approach further through voice conversion. Each frame in T can be decomposed and treated in a similar manner as described in Section 2.1.

The short interval or sequence of frames in T can be represented as:

$$S_T = \{S_{t_1}(f), S_{t_2}(f), ..., S_{t_n}(f)\}, \text{ and} \quad (6)$$

$$H_T = \{H_{t_1}(f), H_{t_2}(f), ..., H_{t_n}(f)\} \quad (7)$$

Each frame $t_i \in T$ can be reconstructed from their corresponding elements in S_T and H_T. While S_T captures the excitation source, which has little influence on ASV, H_T captures the vocal tract response from which cepstral features are extracted. Since it has no impact on ASV the phase information in (7) is discarded in practice.

We aim to estimate a new set of transfer functions F_T to replace H_T in (7) in order to synthesise a new artificial signal more likely to spoof the ASV system, and consequently a more stringent test of vulnerabilities. In the same way as in (5), F_T can be split into gains G_t and frequency responses $A_t(f)$ giving sequences:

$$G_T = \{G_{t_1}, G_{t_2}, ..., G_{t_n}\} \quad (8)$$

$$A_T = \{A_{t_1}(f), A_{t_2}(f), ..., A_{t_n}(f)\} \quad (9)$$

where each $A_t(f)$ is obtained from p prediction coefficients for frame t, i.e. $P_t = \{a_{it}\}_{i=1}^{p}$. The prediction coefficients for the sequence are denoted by P_T.

$$P_T = \{P_{t_1}, P_{t_2}, ..., P_{t_n}\} \quad (10)$$

We then seek a set of parameters to synthesize a new signal which maximises the ASV score according to the following objective function:

$$(P_T^*, G_T^*) = \arg \max_{P_T, G_T} l(f(P_T, G_T, S_T), \lambda_X, \lambda_{UBM}) \quad (11)$$

where, $f()$ is a function which reconstructs a signal from the parameters G_T, P_T and S_X, and $l()$ is the ASV function that scores the generated signal with respect to the target speaker model λ_X and the universal background model λ_{UBM}. Note that the ASV system has a dual role both in identifying the short interval T and in the subsequent optimisation process.

The $(p+1) * n$ variables comprising the prediction coefficients, gains and ASV score in (11) are continuous valued and the optimization problem is non-convex and possibly discontinuous in nature. In our work (11) is maximised with a genetic optimisation algorithm. Genetic algorithms are well-suited to the stochastic nature of the speech signals and have been applied previously in related work, e.g. voice conversion [11, 12] and speech synthesis [13].

A schematic representation of the optimization problem is illustrated in Figure 1. The target speaker spoofing utterance X is used to learn the target speaker model λ_X as well as in the selection of the short segment T for constructing the artificial signals. The dashed block represents the optimisation objective function.

3. EXPERIMENTAL WORK

This section reports our experimental work to test the vulnerabilities of two ASV systems to spoofing from artificial signals.

3.1. ASV systems

The two ASV systems are based on the LIA-SpkDet toolkit and the ALIZE library [14] and are directly derived from the work in [15].

Both systems use a common parametrisation where features are composed of 16 linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy. A common energy-based speech activity detection (SAD) system is also used to remove non-speech frames.

The first ASV system is a standard GMM system with a universal background model (GMM-UBM). The second ASV system includes channel compensation based on factor analysis (FA), with the symmetrical approach presented in [16]. The background data for the channel matrix comes from the NIST Speaker Recognition Evaluation (SRE)'04 dataset. That used for UBM learning comes either from the NIST SRE'04 or NIST SRE'08 datasets depending on whether the ASV system is used to build artificial signals or to assess vulnerabilities to spoofing. This is discussed further in Section 3.3.

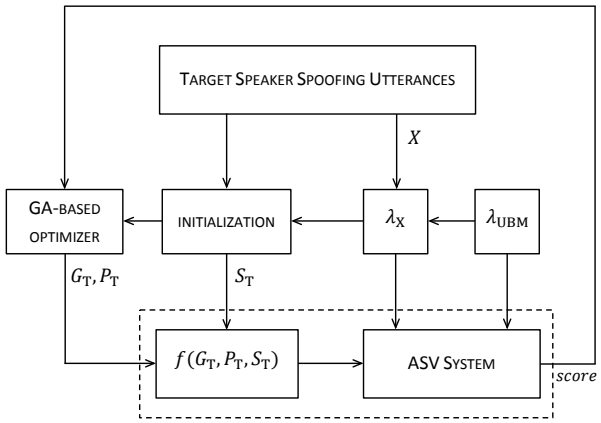


Fig. 1: Schematic representation of the optimization loop

3.2. Experimental protocol

All experiments use data from the 8conv4w-1conv4w task of the NIST SRE'05 dataset. Only one of the 8 training conversations is ever used for training whereas the other 7 are used to build artificial spoofing signals. One conversation provides an average of 2.5 minutes of speech (one side of a 5 minute conversation).

We used only the male subset which contains 201 client speakers. Baseline experiments use the standard NIST test data which results in 894 true client tests and 8962 impostor test. In all spoofing experiments, the number of true client tests is the same as for the baseline whereas the number of impostor tests come from 201 artificial signals (one for each speaker model) resulting in 201 impostor tests.

3.3. Artificial signal generation

Artificial signals are generated as illustrated in Figure 1. The ASV system is the GMM-UBM presented in Section 3.1. The speech signal X is divided into frames of 20ms with a frame overlap of 10ms. ASV scores are generated for each frame in order to identify the short interval $T = \{t_1, \ldots, t_n\}$ in X with the highest average score. We conducted experiments with values of n between 1 and 20 frames and observed good results with a value of $n = 5$. G_T and A_T in (8) and (9) are then calculated in the same way as described in Section 2.1 and in [7].

The genetic algorithm was implemented using the MATLAB Global Optimization Toolbox V3.3.1. Except for the maximum number of generations which is set to 50, we used MATLAB's default configuration. The initial population is composed of all possible combination of the 6 highest scored frames in X thereby producing a population of $\frac{6!}{(6-5)!} = 720$ samples. The lower and upper bounds are formed by adding a margin of 50% to the extremal values of each component in

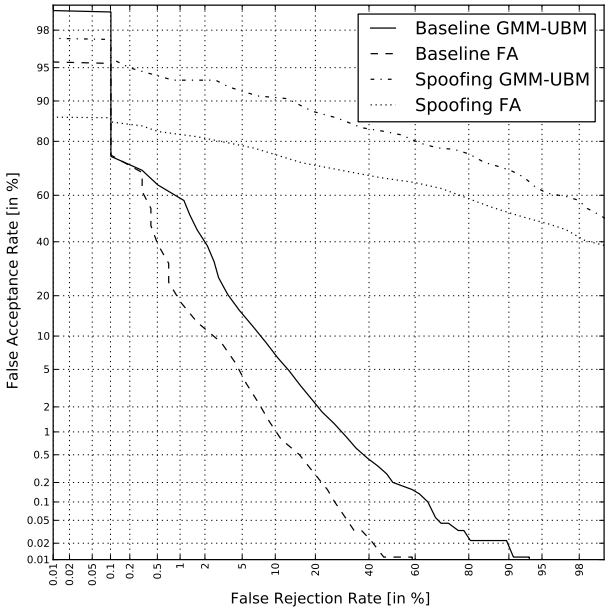


Fig. 2: DET plots for the GMM-UBM and FA ASV systems with and without spoofing through artificial signals.

the initial population. Constraints are related to the stability of the vocal track filter, i.e. the reflection coefficients of each $P_{t_i} \in P_T$ must be all positive.

Finally, we note that the ASV system used to generate artificial signals need not necessarily be the same as that targetted by spoofing. In our experiments the GMM-UBM system is used in the optimisation problem described above with a UBM trained on NIST SRE'08 data. The GMM-UBM and FA ASV systems presented above use a UBM trained on NIST SRE'04 data.

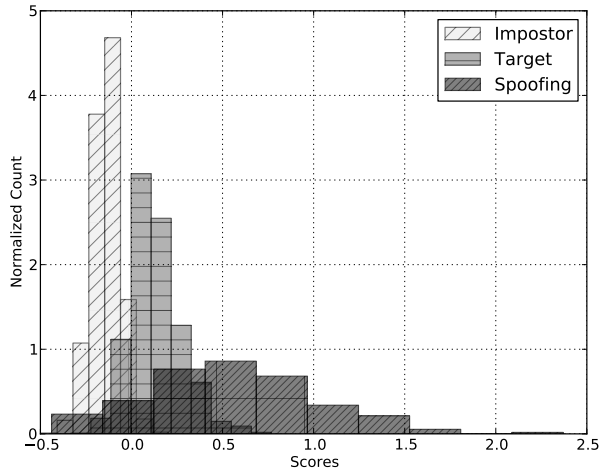
3.4. Results

ScoreToolKit is a TABULA RASA tool for analysing biometric system performance

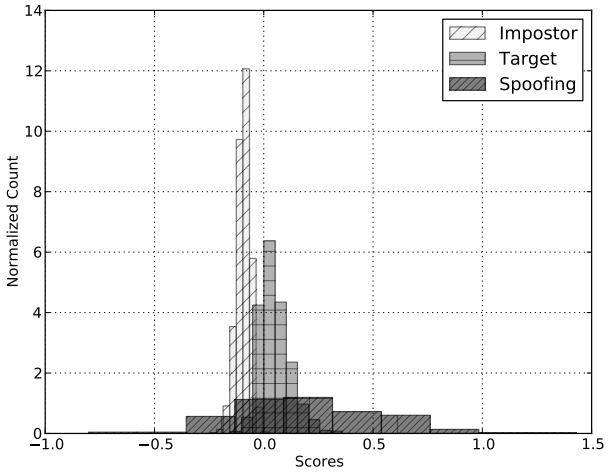
Figure 2 shows detection error trade-off (DET) plots[2] for the two ASV systems with and without spoofing with artificial signals. They show that an equal error rate (EER) of 8.5% for the basic GMM-UBM system rises to 77.3% when all impostor trials are replaced with artificial spoofing signals. The degradation in performance is less pronounced for the FA system for which the EER increases from 4.8% to 64.3%.

In order to compare the impact of artificial signals to that of voice conversion, experiments with the latter were conducted using the same experimental protocol but with converted voice signals generated according to [7]. EERs of 32.6% and 24.8% were obtained for the GMM-UBM and FA systems respectively. These results show that artificial signals potentially pose a comparatively greater threat than voice conversion.

[2]TABULA RASA score toolkit (`http://publications.idiap.ch/downloads/reports/2012/Anjos_Idiap-Com-02-2012.pdf`)



(a) GMM-UBM system



(b) FA system

Fig. 3: Score distributions for target, impostor and spoofing tests with the GMM-UBM and FA systems.

Figures 3(a) and 3(b) illustrate score histograms for target, impostor and spoofing tests for the GMM-UBM and FA ASV systems respectively. In both cases, while the impostor distributions (no spoofing) lie to the left of the target distributions, the spoofing distributions lie to the right, i.e. scores from spoofing tests are greater than for true client tests. The score distributions in Figures 3(a) and 3(b) thus show that artificial spoofing signals are extremely effective in provoking high ASV scores and account for the significant increases in EER observed in Figure 2.

4. CONCLUSIONS AND FUTURE WORK

This work assesses the vulnerability of text-independent automatic speaker verification systems to spoofing with novel artificial, tone-like signals. They are shown to provide an order of magnitude increase in the baseline EER of two different ASV systems.

Even if it not strictly the case with reported experiments, the approach to spoofing investigated in this paper requires no prior knowledge of the targetted system. Furthermore, while the experiments reported in this paper used relatively large quantities of speaker-specific training data to learn speaker models used to optimise spoofing attacks, there is a high probability that models trained on considerably fewer data will also work well so long as they are trained on a similar quantity of data as that used in ASV enrollment.

Having the potential to pass both energy-based and pitch-based voice activity detection systems, artificial signals thus pose a serious threat to the reliability of ASV systems. In line with previous, related work, this contribution highlights the importance of efforts to develop dedicated countermeasures, some of them trivial, to protect ASV systems from spoofing. A straightforward speech quality assessment routine, for example, may be used to distinguish artificial signals from genuine speech signals. The development of such countermeasures, the influence of the quantity of speech data used to generate artificial signals and the effect of such signals when injected into a real impostor speech signal are all subjects of our future work.

5. REFERENCES

[1] M. Faundez-Zanuy, "On the vulnerability of biometric security systems," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 6, pp. 3 – 8, june 2004.

[2] M. Farrs, M. Wagner, J. Anguita, and J. Hern, "How vulnerable are prosodic features to professional imitators?," in *Odyssey*, 2008.

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

[4] 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.

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

[6] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP : Indexation in a Client Memory," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 1, pp. 17 – 20.

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

[8] 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.

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

[10] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *IEEE International Conference on Acoustics Speech and Signal Processing*, march 2010, pp. 1798 –1801.

[11] G. Zuo and W. Liu, "Genetic algorithm based RBF neural network for voice conversion," in *Fifth World Congress on Intelligent Control and Automation*, june 2004, vol. 5, pp. 4215 – 4218.

[12] C. Zhi and Z. Ling-hua, "Voice conversion based on Genetic Algorithms," in *12th IEEE International Conference on Communication Technology (ICCT)*, nov. 2010, pp. 1407 –1409.

[13] P.-Y. Oudeyer, "The production and recognition of emotions in speech: Features and Algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157 – 183, 2003.

[14] 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.

[15] 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.

[16] 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.