Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals

Federico Alegre, Ravichander Vipperla and Nicholas Evans

Multimedia Communications Department, EURECOM, Sophia Antipolis, France
{alegre,vipperla,evans}@eurecom.fr

Abstract

The vulnerability of automatic speaker verification systems to imposture or spoofing is widely acknowledged. This paper shows that extremely high false alarm rates can be provoked by simple spoofing attacks with artificial, non-speech-like signals and highlights the need for spoofing countermeasures. We show that two new, but trivial countermeasures based on higher-level, dynamic features and voice quality assessment offer varying degrees of protection and that further work is needed to develop more robust spoofing countermeasure mechanisms. Finally, we show that certain classifiers are inherently more robust to such attacks than others which strengthens the case for fused-system approaches to automatic speaker verification.

Index Terms: automatic speaker verification, biometrics, spoofing, imposture, countermeasures

1. Introduction

It is now widely acknowledged that automatic speaker verification (ASV) systems are vulnerable to spoofing attacks. A growing body of work shows that significant increases in false alarms can be provoked through impersonation [1, 2], replay attacks [3, 4], voice conversion [5, 6, 7] and speech synthesis [8, 9]. All of this work assumes that spoofing efficacy depends on the projection of speech with acceptable quality. Given that most state-of-the-art ASV systems do not include any form of speech quality assessment, this assumption is unfounded; our own recent work [10] highlights significant vulnerabilities to entirely artificial, non-speech-like tone signals. Dedicated countermeasures to protect ASR systems from spoofing are thus needed urgently.

This is the objective of the EU FP7 TABULA RASA project[1] which aims to develop new spoofing countermeasures for several different biometric modalities including ASV. Our own work aims to systematically evaluate vulnerabilities and, more importantly, to develop dedicated countermeasures to protect systems from attacks through voice conversion and artificial signals. This paper reports recent work involving the latter and demonstrates the level of spoofing protection afforded through two, trivial countermeasures involving higher-level, dynamic features and speech quality assessment.

The remainder of this paper is organised as follows. Section 2 describes the generation of artificial signals which are used to highlight the vulnerability of ASV systems to spoofing. Section 3 describes spoofing countermeasures. Section 4 reports an assessment of spoofing vulnerabilities and of countermeasure protection. Finally, our conclusions are presented in Section 5.

2. Spoofing with artificial signals

Our approach to assess the vulnerability of ASV systems using artificial signals is presented in [10]. It is based on the modification of the voice conversion algorithm reported by other authors [6]. Both are summarised briefly in the following.

2.1. Voice conversion

Bonastre *et al.* [6] show how a speech signal Y represented in the spectral domain according to the standard source-filter model:

$$Y(f) = H_y(f)S_y(f) \tag{1}$$

can be mapped toward a target speaker signal X by replacing $H_y(f)$ in Equation 1 with $H_x(f)$, where $H_{x/y}(f)$ is the vocal tract transfer function of X/Y and $S_{x/y}(f)$ is the Fourier transform of the excitation source. $H_x(f)$ can be estimated using two parallel sets of Gaussian mixture models (GMM) of the target speaker. If the phase of the impostor signal is left unaltered, Y is thus mapped toward X in the spectral-slope sense which is sufficient to overcome most ASV systems. Full details can be found in [6].

2.2. Artificial signal generation

Certain short intervals in a speech signal X, e.g. those corresponding to voiced regions, give rise to higher scores or likelihoods than others and the chances of a spoofing attack succeeding can thus be increased by concentrating on a short interval or sequence of frames in $X = \{x_1, ..., x_m\}$ which gives rise to the highest score.

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 modeled. In order to produce a sample 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 concate-

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

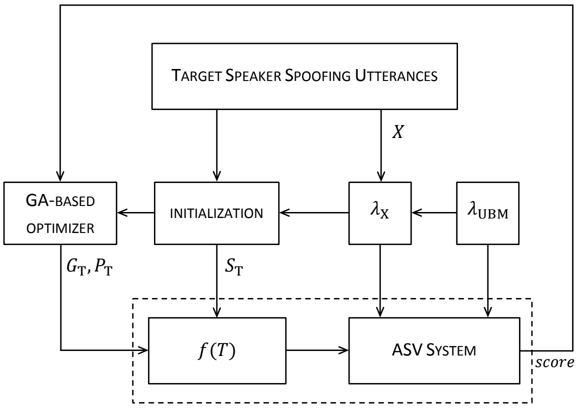

Figure 1: *Optimization loop for artificial signal generation, reproduced from [10].*

nated 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. Based on an initial segment T, enhanced, artificial signals which maximise the potential for spoofing are generated by searching for an optimal sequence of speech frames T^* such that the final score of a given ASV system is maximized. To find T^*, we use an optimization loop as illustrated in Figure 1.

According to Equation 1 the short interval or sequence of frames in T can be represented as:

$$S_T = \{S_{t_1}(f), S_{t_2}(f), ..., S_{t_n}(f)\}, \text{ and} \quad (2)$$

$$H_T = \{H_{t_1}(f), H_{t_2}(f), ..., H_{t_n}(f)\} \quad (3)$$

Each frame $t_i \in T$ can be reconstructed from their corresponding elements in S_T and H_T. Therefore, each frame $t_i \in T$ is transformed in the same manner as that for voice conversion, replacing the corresponding $H_{t_i} \in H_T$. S_T is kept fixed during the optimization process while a genetic algorithm is used for searching optimal parameters of H_T such that the reconstructed speech $f(H_T, S_T)$ provokes a high ASV score. Full details are presented in [10].

3. Spoofing countermeasures

Some state-of-the-art ASV parameterisations and systems capture and utilise speech characteristics at the utterance level. For example, ASV systems based on GMM supervectors inherently capture speech variability and we thus hypothesize that such systems will be naturally robust to spoofing attacks with artificial signals. This hypothesis is investigated in our experiments reported in Section 4. Here we describe two spoofing countermeasures which are independent of recognition and are thus applicable to any ASV system. The first is based on the use of longer contexts, or higher level features, and the second on voice quality assessment.

3.1. Higher level features

Features extracted over a longer context have utility in many applications of speech processing, for example prosodic features extracted over longer contexts have been applied successfully in speaker verification and emotion recognition. Given a speech utterance, longer context higher level features can be extracted at the frame level, word level, phrase level or at the utterance level. In this work, we investigate an utterance level feature that can be quickly computed from frame level features. As described in Section 2 it is very robust in distinguishing between real speech and artificial spoofing signals.

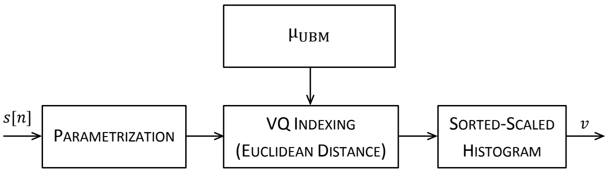

Figure 2: *Block diagram of utterance-level feature (v) generation.*

Figure 2 shows a block diagram of our approach to compute utterance level features. Conventional parameters extracted from the input signal are indexed using vector quantization (VQ) with respect to the means of the UBM that act as VQ centroids. The histogram of the resulting index vector is reordered based on the occurrence frequencies as in a Pareto chart and the frequencies are scaled with respect to the first component to obtain a new feature vector v.

During the vector quantization step, parameters corresponding to a real speech signal are expected to be uniformly indexed to all or most of the Gaussian components, while a tone-like signal with a repetitive pattern is more likely to be associated with one or very few components. Hence for a real speech utterance, the feature v will have a smoother exponential distribution while tone-like artificial signals will exhibit a Dirac, delta-like distribution with a dominant peak in the first coefficient. This facilitates robust classification between real speech and artificial spoofing signals. We used a simple thresholded cosine distance between the test feature vector v_{test} and a mean feature vector v_{mean} estimated from real speech training utterances.

3.2. Voice quality assessment

An intuitive approach to counter spoofing with artificial signals involves voice quality assessment. While a front-end automatic speech recognition (ASR) system could be used as a means of voice quality/intelligibility assessment it is comparatively more complex than the simple approach proposed here. We use the state-of-the-art, standard ITU-T P.563 recommendation [11] voice quality assessment tool. As a single-ended algorithm it forms an ideal countermeasure to identify artificial spoofing signals. The P.563 tool calculates a Mean Opinion Score (MOS) for a given utterance which indicates subjective quality. The MOS is in the range of 1 (worst) to 5 (best). We note that the P.563 tool has already been used in NIST speaker recognition evaluations to estimate speech quality [12] and in the quality assessment of synthesized speech [13, 14]. To the best of our knowledge, there is no previous work in the use of such assessments for spoofing countermeasures.

4. Experimental work

This section reports our experimental work to test vulnerabilities to spoofing from artificial signals, with and without spoofing countermeasures.

4.1. ASV systems

Experiments were conducted with five different automatic speaker verification (ASV) systems using combinations of different parameterisations and classifiers. All systems are based on the LIA-SpkDet toolkit and the ALIZE library [15] and are directly derived from the work in [16].

The first parameterisation comprises 16 linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy resulting in a feature vector of 33 coefficients. A second parameterisation is used to assess performance when the ASV system used for spoofing is different to that being assessed. It is the same as the first parameterisation except with appended second derivatives and delta-delta energy thereby giving features vectors of 50 coefficients. The two parameterisation are hereafter referred to as 'm33' and 'm50' respectively. In both cases the same energy-based speech activity detection (SAD) system is applied to remove non-speech frames.

The first ASV system is a standard Gaussian mixture model (GMM) system with a universal background model (UBM). The second ASV system includes channel compensation based on factor analysis (FA) with the symmetrical approach presented in [17]. The third system is a support vector machine (SVM) classifier applied to GMM supervectors which come directly from the GMM-UBM system.

4.2. Experimental protocol

ASV experiments relate to the 8conv4w-1conv4w task of the NIST SRE'05 dataset. One of the 8 training conversations per speaker is used for training and another, different conversation is used for generating artificial signals. Each conversation has an average duration of 2.5 minutes.

Results are reported for the male subset only and, for spoofing assessments, all impostor tests in the baseline corpus are replaced with artificial signals for which there is only one per speaker.

Background data used for UBM learning and channel modelling comes from the NIST SRE'04 dataset. That used to generate artificial signals comes from the NIST SRE'08 dataset. Whereas ASV assessments consider both parameterisations and all three classifiers, the ASV system used to generate artificial signals uses only parameterisations of 33 coefficients and the baseline GMM-UBM classifier.

4.3. Baseline and spoofing results

Baseline results for each parameterisation/classifier combination are illustrated in terms of equal error rates (EERs) in Table 1. For the same GMM-UBM system and m33 parameterisation used to generate artificial signals, results show a marked degradation in performance from an EER of 8.5% to 77.1%. While the baseline performance for the FA is significantly better, a similar degradation is observed under spoofing (4.8% to 64.2%).

System	Baseline	Spoofing
GMM_m33	8.5	77.1
FA_m33	4.8	64.2
GMM_m50	7.7	66.3
FA_m50	4.2	57.7
SVM_m33	7.8	4.1

Table 1: *EERs (%) for baseline system and under spoofing.*

With m50 parameterisations baseline performances are slightly improved in both cases and the difference to parameterisations used to generate artificial signals affords a slightly improved robustness to spoofing, however, the EERs in both cases still remain high (66.3% and 57.7%).

Finally, as expected, the SVM approach shows significantly better robustness. While the baseline EER of 7.8% is worse than that of the FA system, almost all spoofed tests are detected and the EER falls to 4.1%.

4.4. Countermeasures

Both the high level feature (HLF) and voice quality assessment (p563) countermeasures operate independently of verification and act as a filter to differentiate real speech utterances from artificial signals. Being a similar two class problem, results are reported in terms of EERs as before and using same spoofing corpus used for ASV experiments.

The ITU standard p563 voice quality assessment tool can be applied to speech signals of between 3 and 20 seconds with a speech activity ratio (SAR) in the range of 25 to 75%. Since all utterances used here are in the order of 2.5 minutes in duration, they are split into segments of 15 seconds. The SAR of each segment is estimated using the ASV energy-based speech detector and the p563 tool is applied to all those segments with a satisfactory SAR. Results in terms of score distributions and spoofing detection performance are illustrated[2] in Figure 3 and Table 2 respectively. Figure 3 shows that artificial signals produce a large number of scores below 1.5 but also a small number of higher scores. Real speech signals produce a balanced spread with a mean of approximately 2.5. There is thus some potential to detect artificial signals though the two distributions do overlap. The p563 countermeasure is moderately successful in preventing some artificial signals in being passed to the ASV system but also leads to some false alarms, i.e. valid speech signals withheld from ASV.

HLF performance was assessed in an identical fashion. We observed that the distributions for real and artificial signals do not overlap and thus a perfect separation is possible. Table 2 thus shows an EER of 0% for the HLF countermeasure. The performance of the p563 countermeasure is rather poor, but it is based only on prior knowledge of speech, and not on the particular spoofing attack investigated. While the HLF approach does give a perfect separation of real speech and artificial signals it is rather specific to the artificial spoofing signals investigated. It may thus be overcome by alternatively generated artificial sig-

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

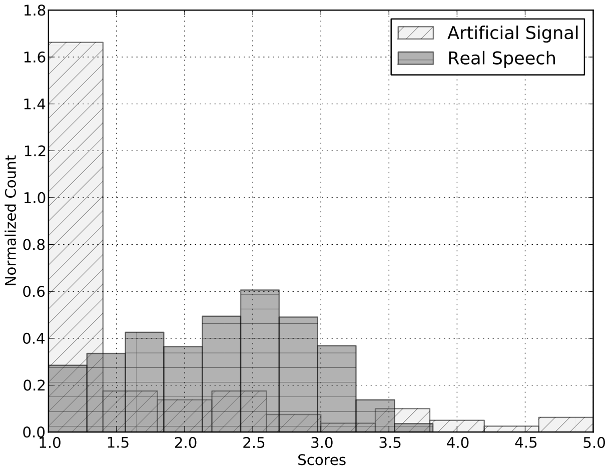

Figure 3: *Performance of ITU-P563 as attack detector.*

System	EER
p563	27
HLF	0

Table 2: *Spoofing detection performance in terms of EER (%)*

nals, whereas p563 may be more robust. In this sense, even though the p563 countermeasure generates a significant number of false alarms, it is not without merit.

5. Conclusions and future work

This work assesses the vulnerability of text-independent automatic speaker verification (ASV) systems to spoofing with artificial signals. Large increases in EER are reported for two popular approaches to ASV and the use of different parameterisations does not afford any significant protection. Trivial countermeasures based on speech quality assessment and high level features (HLFs) offer varying levels of protection for varying degrees of generality with respect to spoofing approaches. Finally, we show that a support vector machine (SVM) classifier is inherently robust to spoofing attacks with artificial signals and strengthens the case for fused-system approaches to automatic speaker recognition.

While SVM approaches can behave well on unseen data, and despite the perfect performance obtained with HLFs, it is likely that they too can be overcome using optimisation approaches similar to those used in generating artificial signals. While simple countermeasure solutions may be effective against known attacks, more sophisticated and general countermeasures solutions are needed to ensure robustness to unforeseen forms of spoofing. One promising direction for future work involves the use of frame-level score distributions rather than averaged frame scores to better protect ASV systems from spoofing with artificial signals.

6. References

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

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

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] L. Malfait, J. Berger, and M. Kastner, "P.563–The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924 –1934, nov. 2006.

[12] A. Harriero, D. Ramos, J. Gonzalez-Rodriguez, and J. Fierrez, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *Advances in Biometrics*, vol. 5558 of *Lecture Notes in Computer Science*, pp. 434–442. Springer Berlin / Heidelberg, 2009.

[13] S. Möller, "Estimating the quality of transmitted synthesized speech with the single-ended quality prediction model according to ITU-T Rec. P.563," ITU-T Study Group 12 - Delayed Contribution 174, Geneva, Switzerland, 2006.

[14] I. Kraljevski, S. Chungurski, I. Stojanovic, and S. Arsenovski, "Synthesized speech quality evaluation using ITU-T P.563," 18-th TELFOR, Belgrade, Serbia, November 2010.

[15] 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.

[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] 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.