SPOOFING COUNTERMEASURES TO PROTECT AUTOMATIC SPEAKER VERIFICATION FROM VOICE CONVERSION

Federico Alegre, Asmaa Amehraye and Nicholas Evans

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

ABSTRACT

This paper presents a new countermeasure for the protection of automatic speaker verification systems from spoofed, converted voice signals. The new countermeasure exploits the common shift applied to the spectral slope of consecutive speech frames involved in the mapping of a spoofer's voice signal towards a statistical model of a given target. While the countermeasure exploits prior knowledge of the attack in an admittedly unrealistic sense, it is shown to detect almost all spoofed signals which otherwise provoke significant increases in false acceptance. The work also discusses the need for formal evaluations to develop new countermeasures which are less reliant on prior knowledge.

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

1. INTRODUCTION

It is widely acknowledged that automatic speaker verification (ASV) systems are vulnerable to spoofing. While earlier work considered the threat from classical spoofing attacks such as impersonation [1,2] or replay [3,4], that from more advanced attacks has attracted more recent attention. Over the last six years attacks from voice conversion [5–8], speech synthesis [9, 10] and artificial signals [11] have all been shown to provoke significant increases in the false acceptance rate of state-of-the-art ASV systems.

Reassuringly, as has been the case for other biometric modalities, e.g. face recognition [12–14], the speaker recognition community has started to address the problem through efforts to develop specific spoofing countermeasures [15–20]. Prior work to develop spoofing countermeasures has focused mostly on those aspects of the speech signal not used for recognition. Several works have demonstrated increased robustness to spoofing stemming from speech synthesis through the use of prosodic features and phase [15–17]. Based on the assumption that phase information is generally lost during voice conversion, the utility of phase-related features to detect voice conversion was investigated in [18,19]. While both approaches are effective in reducing false acceptance rates, spoofing is however far from being a solved problem.

Our work in the EU Tabula Rasa project[1], which aims to develop spoofing countermeasures for a number of different biometric modalities, has shown voice conversion to be particularly difficult to detect. This work considered an approach to voice conversion originally proposed in [7]. It shows the limitations of ASV systems when only the spectral slope is altered. As such, converted speech retains real-speech phase information. Attacks of this nature will thus have the potential to overcome the countermeasures proposed in [15–20]. This paper reports a new countermeasure which exploits the reduction in pair-wise distances between consecutive feature vectors when they are both shifted towards the same local maxima of the likelihood function of a target speaker model as a consequence of voice conversion. We accept that the proposed countermeasure exploits prior knowledge of the spoofing attack and discuss this issue later in the paper.

The remainder of this paper is organized as follows. The voice conversion approach and new countermeasure are presented in Sections 2 and 3, respectively. Experimental work is described in Section 4. Finally our conclusions and ideas for future work are presented in Section 5.

2. VOICE CONVERSION

All work reported in this paper was conducted with our implementation of the approach to voice conversion originally proposed in [7]. 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

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

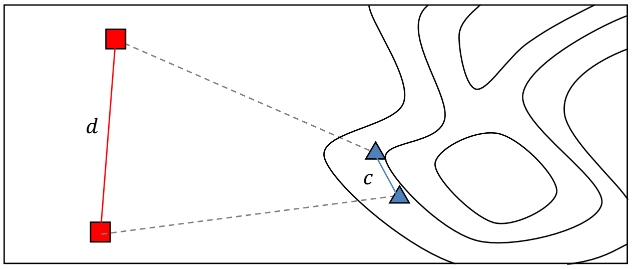


Fig. 1: An illustration of voice conversion in feature space showing the shift of two consecutive feature vectors towards a common local maxima. We usually expect $c < d$.

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 in a spectral-slope sense, which 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:

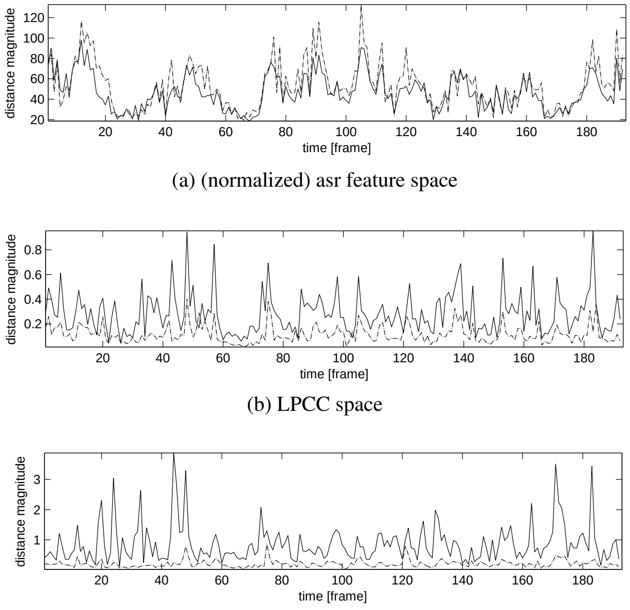
$$x_{fil} = \sum_{i=1}^{M} p(g^i_{asr}|y_{asr})\mu^i_{fil} \qquad (2)$$

where $p(g^i_{asr}|y_{asr})$ is the a posteriori probability of the Gaussian component i given the frame y_{asr} and μ^i_{fil} is the mean of the Gaussian g^i_{fil} tied to the Gaussian g^i_{asr}. $H_x(f)$ is estimated from x_{fil} using an LPCC-to-LPC transformation and a time-domain signal is synthetized from converted frames with a standard overlap-add technique. Full details can be found in [7,21,22].

From Equation 1 we note that this approach to voice conversion retains real-speech phase and excitation. In consequence, the approaches to detect converted voice and synthesized speech reported in [15–20] will not detect speech converted according to the approach described above.

3. SPOOFING COUNTERMEASURE

The work presented in this paper is conducted with full prior knowledge of the spoofing attack. We accept that this is wholly unrealistic of a practical spoofing scenario but must be content with such an approach. Future evaluations with independent efforts in spoofing and countermeasures will facilitate the development of the latter without prior knowledge of the spoofing attack(s).



(a) (normalized) asr feature space

(b) LPCC space

(c) LPC space

Fig. 2: *An illustration of the pair-wise distance between consecutive feature vectors for asr, LPCC and LPC parameterisations. Profile shown for genuine speech (solid line profiles) and converted voice (dashed line profiles).*

3.1. Features

Voice conversion shifts the spectral slope of a spoofer towards that of a target, according to Equation 2. Note that if the term μ^i_{fil} is replaced by μ^i_{asr}, then Equation 2 is identical to the expectation step in the Expectation Maximization (EM) algorithm [23].

The principal behind our countermeasure exploits the expected shift of consecutive feature frames towards the same, closest local maxima of the likelihood function of a particular target model. This principal is illustrated in Figure 1 for two consecutive feature vectors in a two-dimensional space. Under such conditions the relative distance between consecutive feature vectors (red squares) will be reduced (blue triangles) whereas the density of features surrounding the local maxima will be increased. Accordingly we have investigated a new countermeasure to detect this phenomenon in order to distinguish between converted voice and a genuine speech signal.

We conducted initial experiments independently from ASV to validate this phenomenon. Figure 2 shows plots of the $n - 1$ consecutive, pairwise distances for n frames of example genuine speech and converted voice signals. Plots are illustrated for the same feature sets discussed in Section 2. As shown in Figure 2(a), the differences between genuine speech and converted voice is relatively low in the (normalized) asr feature space; the two profiles are more or less identical. For LPCC space (b), the differences are more significant and in LPC space (c) they are particularly pronounced. Since LPC

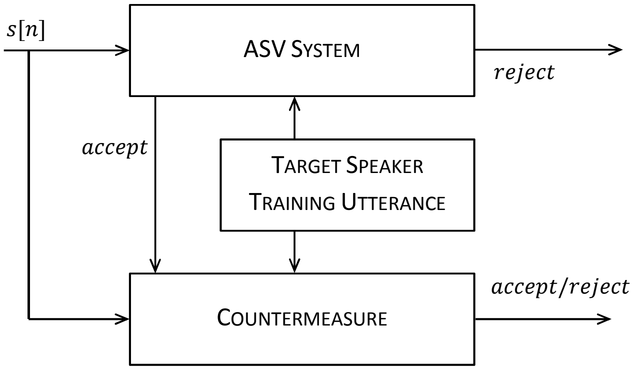


Fig. 3: *A block diagram of the integrated ASV system and proposed countermeasure.*

pair-wise distances exhibit the greatest differences between genuine speech and converted voice, they are used in all subsequent experiments reported here.

Finally, even though we predict increases in cluster density as a result of voice conversion, initial experiments to observe changes in variance were discouraging and thus the use of variance estimates was not pursued further.

3.2. Detection

A block diagram of the integrated ASV system and proposed countermeasure is illustrated in Figure 3. The countermeasure is speaker-dependent and exploits differences in the distribution of pairwise distances between test data $s[n]$ and that used to train the target model in question. The percentage overlap between the two distributions forms a score which is then thresholded to classify $s[n]$ as either genuine speech or converted voice. When the two distributions are normalised, the percentage overlap lies between zero and unity. Lower scores indicate genuine speech whereas higher scores indicate converted voice.

As in other prior work [10,19], and illustrated in Figure 3, the proposed countermeasure is integrated with the ASV system as an independent post processing step. Claimed identities are thus only accepted if a test signal $s[n]$ attains a likelihood higher than the ASV threshold and a countermeasure score lower than its threshold. As discussed later, the combination of two independent classifiers makes assessment somewhat troublesome.

4. EXPERIMENTAL WORK

Here we report experiments to assess the performance of the new countermeasure.

4.1. ASV systems

Experiments were conducted with five ASV systems used within the EU Tabula Rasa project. They are all based on the LIA-SpkDet toolkit [24] and the ALIZE library [25] and are directly derived from the work in [26]. In all cases the speech signal is divided into frames of 20ms with a frame overlap of 10ms. All systems use a common parameterization where features extracted using SPro [27] 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 and all systems use a common universal background model (UBM) with 1024 Gaussian components.

The first ASV system is a standard Gaussian mixture model (GMM) system with a UBM denoted (GMM-UBM). The second system includes channel compensation based on factor analysis (FA) according to the symmetrical approach presented in [28]. The third system is a support vector machine (SVM) classifier which is applied to GMM supervectors coming directly from the GMM-UBM system. It is referred to as a GMM supervector linear kernel system (GSL). The fourth system is almost identical to the third but is enhanced with nuisance attribute projection [29] to attenuate intersession (interchannel) variability, with NAP matrices of rank 40. The fifth approach is a GSL system with FA supervectors (GSL-FA) [26].

4.2. Protocols and metrics

Speech data used for UBM learning comes either from the NIST Speaker Recognition Evaluation 2004 (NIST‘04) or NIST‘08 datasets depending on whether the resulting GMM is used for the baseline ASV systems or for the conversion system respectively. Background data used for nuisance attribute projection (NAP) and factor analysis (FA) comes from the NIST‘04 dataset.

The male subset of the NIST‘05 dataset is used for development of both ASV systems and the countermeasure, whereas the NIST‘06 dataset is used for evaluation. As in [20], 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 countermeasures 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.

Standard NIST protocols dictate in the order of 1000 true client tests and in the order of 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 is converted toward the corresponding client model. Finally, given the consideration of spoofing and without any specific, standard operating criteria, 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.

Error	EER (%)		FAR (%)	
System	Baseline	Spoofing	Baseline	Spoofing
GMM-UBM	9	32.6	6	77
GSL	8.5	37.2	6	88
GSL-NAP	7	32.1	3	84
FA	5.5	24.4	1	54
GSL-FA	6.5	30.3	2	82

Table 1: *ASV performance in terms of EER and FAR for a fixed FRR of 10%.*

4.3. Voice conversion system and countermeasure setup

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. The front end processing used in voice conversion is thus exactly the same as that used for ASV. The filtering model and filter $H_x(f)$ uses 19 LPCC and LPC coefficients respectively.

The countermeasure operates on the same 19^{th} order LPC vectors recalculated from a time domain signal $s[n]$ in Figure 3. Frame blocking is the same as for ASV systems and voice conversion (although different frame lengths do provide similar results). We take into account only those frames determined to contain voiced speech. Voiced speech was detected using the robust algorithm for pitch tracking (RAPT) [30] in the VOICEBOX toolkit[2] with a default configuration.

4.4. Results

We report the effect of voice conversion on the five ASV systems considered and then the performance of the proposed countermeasure to detect converted voice.

A summary of verification performance in terms of EER for the evaluation set is illustrated in Table 1 for all five ASV systems. Also shown are false acceptance rates (FARs) for a corresponding false rejection rate of 10%. Results show that the EERs of all five ASV systems increase significantly when impostor tests are replaced with converted voice. Only the FA system shows an EER of less than 30% whereas that for the GSL system increases to almost 40%. The scale of the threat is perhaps best illustrated in terms of FAR which increases from 6% to 88% for the GSL system and from 1% to 54% for the FA system.

Ideally, we would report similar EER and FAR statistics for the full system when the countermeasure is integrated with ASV. This is a non-trivial problem, however, for which standard evaluation protocols have yet to be defined. While this work is underway in the EU Tabula Rasa project, before it is complete we prefer to assess countermeasures independently from ASV.

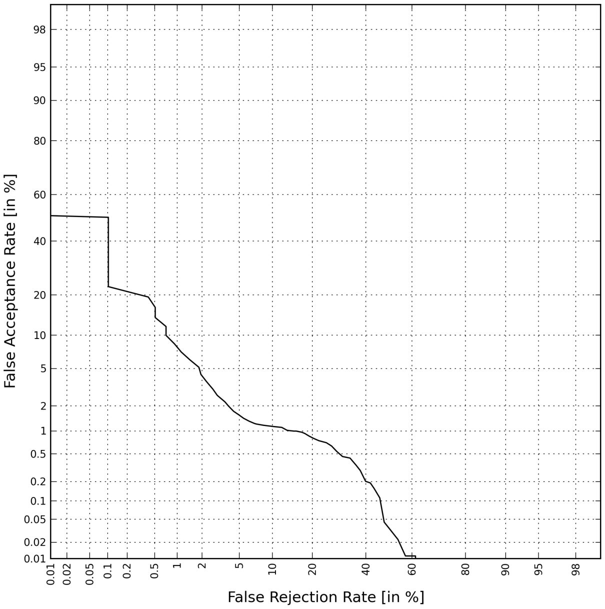


Fig. 4: *A DET profile illustrating countermeasure performance*

Countermeasure performance is illustrated with a DET plot in Figure 4. It illustrates performance as a two-class problem to distinguish genuine speech from converted voice. The new countermeasure is shown to perform exceptionally well and delivers an EER of 2.7%.

5. CONCLUSIONS AND FUTURE WORK

This paper investigates the protection of automatic speaker verification (ASV) systems from spoofing with converted voice. While five tested ASV systems show considerable vulnerabilities, the new countermeasure is shown to be consistent and extremely effective in detecting spoofed speech signals.

While the work reported in this paper is successful in overcoming the specific attack considered, in reality system designers and countermeasure developers cannot assume such prior knowledge. Thus, of far greater significance to the community, is the need for formal evaluations to stimulate the research of spoofing countermeasures in a setting where the exact nature of spoofing attacks is unknown and varied. The development of effective countermeasures will then be extremely challenging.

[2] `http://http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

6. REFERENCES

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

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

[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] 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.

[6] 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.

[7] 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.

[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 *Proc. EUROSPEECH*, 1999.

[10] 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.

[11] 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.

[12] J. Määttä, A. Hadid, and M. Pietikäinen, “Face spoofing detection from single images using texture and local shape analysis,” *Biometrics, IET*, vol. 1, no. 1, pp. 3–10, 2012.

[13] M.M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, et al., “Competition on counter measures to 2-d facial spoofing attacks,” in *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.

[14] F. Alegre, X. Zhao, N. Evans, J. Bustard, M. Nixon, A. Hadid, W. Ketchantang, S. Picard, S. Revelin, A. Riera, et al., “Tabula rasa trusted biometrics under spoofing attacks,” .

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19 – 41, 2000.

[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. Odyssey IEEE 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] 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.

[27] G. Gravier, “Spro: speech signal processing toolkit,” *Software available at http://gforge. inria. fr/projects/spro*, 2003.

[28] 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.

[29] 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. ICASSP*, may 2006, vol. 1, p. I.

[30] D. Talkin, “A robust algorithm for pitch tracking (rapt),” *Speech coding and synthesis*, vol. 495, pp. 518, 1995.