Introducing I-Vectors for Joint Anti-spoofing and Speaker Verification

Elie Khoury[†], *Tomi Kinnunen*[*], *Aleksandr Sizov*[*], *Zhizheng Wu*[‡], *Sébastien Marcel*[†]

[†]Idiap Research Institute, Switzerland
[*]School of Computing, University of Eastern Finland, Finland
[‡]School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

Any biometric recognizer is vulnerable to direct spoofing attacks and automatic speaker verification (ASV) is no exception; replay, synthesis and conversion attacks all provoke false acceptances unless countermeasures are used. We focus on voice conversion (VC) attacks. Most existing countermeasures use full knowledge of a particular VC system to detect spoofing. We study a potentially more universal approach involving generative modeling perspective. Specifically, we adopt standard i-vector representation and probabilistic linear discriminant analysis (PLDA) back-end for joint operation of spoofing attack detector and ASV system. As a proof of concept, we study a vocoder-mismatched ASV and VC attack detection approach on the NIST 2006 speaker recognition evaluation corpus. We report stand-alone accuracy of both the ASV and countermeasure systems as well as their combination using score fusion and joint approach. The method holds promise.

Index Terms: speaker recognition, spoofing, voice conversion attack, i-vector, joint verification and anti-spoofing

1. Introduction

Biometric person authentication [1] plays an increasingly important role in border control, crime prevention and personal data security. While the main biometric techniques (e.g. face, voice, fingerprints) can already handle noisy and mismatched sample comparisons robustly, recognizer vulnerability under malicious attacks remains a serious concern. Indeed, any biometric system has several weak links [2], the most accessible ones being sensor- and transmission-level attacks. We focus on automatic speaker verification that can be spoofed by replay, impersonation, speech synthesis and voice conversion techniques (refer to [3] for an overview). Due to its flexibility in direct transformation of speaker characteristics, we focus on voice conversion (VC) [4] attacks.

Several independent studies confirm that VC attacks pose a serious threat to any speaker verification system. Early studies [5, 6, 7, 8] showed this to be the case regarding traditional Gaussian mixture model (GMM) recognizers. Recent studies involving both text-independent [9, 10] and text-dependent [11] recognizers highlight that the problem persists even with modern recognizers, including i-vectors [12]. Interestingly, the quality of the converted voice does not have to be particularly high; even artificial signal attacks [13, 14] involving unintelligible speech can spoof a recognizer. Even if the modern recognizers might provide increased protection [9, 14], their false acceptance typically increases by considerable amount. This is easy to understand, remembering that speaker verification and VC methods use *matched* front- and back-end models, namely, Mel-frequency cepstral features and GMMs.

While the above studies confirm the destructive nature of VC spoofing, much less work exists in designing countermeasures to safeguard recognizers from attacks. We identify two subproblems in designing such countermeasures. Firstly, spoofing attacks should be *detected*; while a speaker verification system produces a speaker similarity score, a VC attack detector should assess whether the test utterance involves an intentional speaker identity transformation or not. Secondly, we must *integrate* the speaker verification and countermeasure opinions coherently — which is usually done by a simple cascade or score fusion. Regarding attack detection, most of the current solutions utilize prior knowledge about the VC technique or the type of artefact traces it leaves to converted speech. To exemplify, [10, 15, 16] uses phase information known to be absent in the used voice coder technique while [17, 18] uses knowledge that dynamic variation in synthetic speech is smaller compared to natural speech. Such countermeasures are necessarily designed to detect a particular attack which, however, can never be exactly known in advance. *Generalized* countermeasures, a recent direction in biometric anti-spoofing research, aim at detecting various types of attacks (e.g. synthesis, replay or VC attacks), for instance by modeling only in-class data or using enhanced features such cepstrogram texture [19].

We study generalized countermeasures, too, but approach the problem from a generative probabilistic modeling perspective. More specifically, for the first time, this work proposes the use of standard i-vector utterance representation [12] and probabilistic linear discriminant analysis (PLDA) back-end [20] for spoofing detection. Given the excellent performance of i-vector approach across various speech problems, it is natural to study its usefulness for spoofing attack detection. The main benefit of doing so is that we can treat speaker verification and spoofing detection problems similarly, leading to simple *joint* modeling approach (Fig. 1(a)), rather than subsystem fusion (Fig. 1(b)) lacking correlation modeling across the main biometric modality and a spoofing detector.

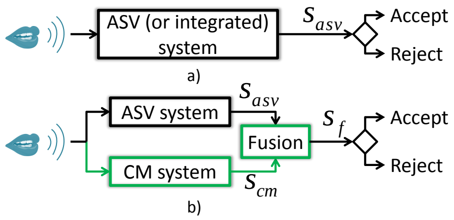

Figure 1: Traditional way to protect automatic speaker verification (ASV) from spoofing attacks is to independently develop ASV and countermeasure (CM) subsystems that are post-combined with score-level fusion (b). Our core contribution is a *joint* approach that uses same i-vectors for both speaker verification and voice conversion attack detection. $s_{\mathbf{x}}$ means a score produced by system $\mathbf{x}$.

Table 1: Statistics of the spoofing dataset used in this work. MCEP and LPC refer to Mel cepstral based VC and linear predictive coding based VC, respectively.

	Male	Female	Total
Target speakers	241	342	583
Genuine trials	1,614	2,332	3,946
Impostor trials	1,132	1,615	2,747
MCEP impostor trials	1,132	1,615	2,747
LPC impostor trials	1,132	1,615	2,747

2. Database and Protocols

Dataset. In this work, we employ the spoofing attack dataset designed in [9, 10]. It is based on the NIST SRE06 corpus, which is a widely used standard benchmark database for text-independent speaker verification. There are 9,440 gender-matched trials for evaluation, consisting of 3,946 genuine trials, 2,747 impostor trials, and 2,747 impostor trials after VC. The voice conversion is implemented by the popular joint-density Gaussian mixture model (JD-GMM) based method [21]. More details of the dataset designing process can be found in [9, 10]. However, unlike previous work where experiments were carried out on only one Mel-cepstral features for VC (MCEP), our study additionally investigates linear predictive coding based features for VC (LPC). The repartition of trials between female and male speakers are reported in Table 1.

Conditions. To study the generalization ability of a countermeasure we define "*matched*" and "*mismatched*" conditions:

- *Matched spoof condition:* This is the most studied case in the literature. It assumes that the user has prior knowledge about the vocoding technique of the VC attacks. For example, if the test set contains trials with MCEP-coded VC, the user may use MCEP-coded synthetic speech while designing his countermeasure. We name this sub-condition "MCEP-MCEP". Similarly, "LPC-LPC" sub-condition is defined.
- *Mismatched spoof condition:* This was usually neglected in previous work. It assumes that the system designer is prepared to a specific type of spoofing, but the attacks are from a different type. In this work, the two related sub-conditions are named "MCEP-LPC" and "LPC-MCEP". For example, MCEP-LPC means that the system is trained to face VC attacks of MCEP, but in fact, the attacks are LPC-coded VC.

In practice, we use the SPTK toolkit[1] to perform MCEP and LPC analysis and synthesis. Similar to [15, 18], a copy-synthesis approach is employed to generate the MCEP- and LPC-coded speech for training the spoofing detector without undergoing any specific VC technique. That is, we first decompose a speech signal into its Mel-cepstral (or LPC) and fundamental frequency (F0) parameters and re-synthesize an approximated signal directly from these parameters. The reconstructed replica, in general, will be close to the original signal but not exactly the same due to the lossy analysis-synthesis model; perceptually, a buzzy or muffled voice quality can be observed. Such copy-synthesis is a straightforward way to generate training samples for spoofing detection without, however, involving the computationally demanding stochastic VC part, which requires selection of source-target speaker pairs and parallel training set. The copy-synthesis speech of SRE04, SRE05 and SRE06 is generated for both MCEP and LPC.

Evaluation Criteria. The evaluation of the ASV system is done in terms of both LICIT and SPOOF protocols [22]. The LICIT protocol, involving *zero-effort* impostors, is the typical evaluation protocol used in verification scenarios, whereas the SPOOF protocol is used to evaluate system performance when spoofing attacks are present. The metrics used for the LICIT protocol are *equal error rate* (EER) and *minimum decision cost function* (mDCF) [23]. The metric used for SPOOF protocol is *spoofing false acceptance rate* SFAR [22] that corresponds to the threshold fixed by the EER on the LICIT protocol. We also plot the detection error trade-off (DET) curves for both protocols. To independently evaluate the countermeasure performance we use *accuracy*, $\mathrm{Acc} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{P}+\mathrm{N})$, where TP is the number of samples correctly classified as positive (i.e. natural speech), TN the number of samples correctly classified as negative (i.e. spoofing attacks), P the total number of positive samples and N the total number of negative samples.

3. I-Vector Extraction

This section briefly describes the complete i-vector [12] extraction and preprocessing chain used in our work. First, a simple energy-based voice activity detection (VAD) is performed to discard the non-speech parts. Second, 19 MFCC and log energy features together with their first- and second-order derivatives are computed over 20 ms Hamming windowed frames every 10 ms. Finally, cepstral mean and variance normalization (CMVN) is applied on the resulting 60-dimensional feature vectors.

The total variability paradigm is built upon Gaussian Mixture Model (GMM) framework [24] and its aim is to extract a low-dimensional vectors, so-called *i-vectors*, that are a compact version of GMM supervectors. Section 5 details the specific parameter values used.

To achieve a higher recognition accuracy we map i-vectors into a more adequate space with the following preprocessing algorithms: (1) *radial gaussianization* [25], which consists of whitening and length-normalization, to reduce non-Gaussian effects as well as mismatch between training and testing subsets, (2) *linear discriminant analysis* (LDA) to learn a linear projection that maximizes between-class variations while minimising within-class variations, (3) *within-class covariance normalization* (WCCN) [12] that normalizes the within-class covariance matrix of training i-vectors.

4. Anti-spoofing and Speaker Verification

4.1. Back-end Speaker Verification

For back-end modeling and scoring we use *probabilistic linear discriminant analysis* (PLDA) [20, 26]. PLDA is a probabilistic framework that incorporates both between- and within-speaker variability, which allows to perform session compensation and to generate log-likelihood ratio (LLR) scores. It assumes that the j-th i-vector $\boldsymbol{\phi}_{i,j}$ of a client i is generated as follows:

$$\boldsymbol{\phi}_{i,j} = \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{i,j} + \boldsymbol{\varepsilon}_{i,j}\ , \qquad (1)$$

where $\mathbf{V}$ and $\mathbf{U}$ are the subspaces describing the between-class and within-class variations respectively. Here $\mathbf{y}_i$ and $\mathbf{x}_{i,j}$ are the associated latent variables, which have a standard normal distribution, $\boldsymbol{\varepsilon}_{i,j}$ represents the residual noise and follow a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with a diagonal covariance matrix. To learn the parameters $\theta = \{\mathbf{V}, \mathbf{U}, \boldsymbol{\Sigma}\}$ of this model we use a scalable EM-algorithm [26] over a training set of i-vectors.

Once the model has been trained, we are able to perform a speaker verification task: for the pair of two i-vectors $(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i)$,

[1] http://sp-tk.sourceforge.net/

where $\boldsymbol{\phi}_t$ is a test i-vector, and $\boldsymbol{\phi}_i$ is an enrolment i-vector of the i-th client, we compute the following LLR score:

$$s_{\text{sv}}(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i) = \log \frac{p(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i|\theta)}{p(\boldsymbol{\phi}_t|\theta)p(\boldsymbol{\phi}_i|\theta)} . \qquad (2)$$

Here, $p(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i|\theta)$ is the probability that the i-vectors $\boldsymbol{\phi}_t$ and $\boldsymbol{\phi}_i$ share the same latent identity variable $\mathbf{y}_i$ and, hence, are coming from the same client, whereas $p(\boldsymbol{\phi}_t|\theta)p(\boldsymbol{\phi}_i|\theta)$ is the probability that the i-vectors $\boldsymbol{\phi}_t$ and $\boldsymbol{\phi}_i$ have different latent identity variables $\mathbf{y}_t$ and $\mathbf{y}_i$ and, therefore, are from different clients. The higher the LLR score the more likely that both i-vectors belong to the same speaker.

4.2. Back-end Spoofing Detection

Spoofing detection is a binary classification task that aims to isolate prepared attacks from natural zero-effort (both genuine and impostor) trials. When dealing with VC attacks, one may look at the problem from a low-level signal processing point of view and solve it by using prior knowledge about the VC technique (e.g. absence of the phase modeling). This was addressed in existing work such as [15, 17].

In this study we present a first attempt to perform the VC detection task applied directly to the i-vectors. We evaluate three different classification methods: fast cosine scoring [12] on average i-vectors, PLDA scoring [20, 26] and *support vector machines* (SVM) [27] with linear kernel.

4.3. Joint Anti-spoofing and Speaker Verification

Score fusion combines scores from multiple systems. We apply it to ASV and Anti-spoofing systems. One of the most powerful score fusion techniques is *logistic regression*, which has been successfully employed for combining heterogeneous speaker classifiers [28, 29].

Let a test i-vector $\boldsymbol{\phi}_t$ be processed by both ASV and countermeasure systems. Each system produces an output score, $s_{\text{asv}}(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i)$ and $s_{\text{cm}}(\boldsymbol{\phi}_t)$ for speaker verification and anti-spoofing (countermeasure) respectively. The final fused score is expressed by the logistic function:

$$s_f(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i|\,\boldsymbol{\beta}) = g\left(\beta_0 + \beta_1 s_{\text{asv}}(\boldsymbol{\phi}_t, \boldsymbol{\phi}_i) + \beta_2 s_{\text{cm}}(\boldsymbol{\phi}_t)\right), \quad (3)$$

where $g(x) = 1/(1 + \exp(-x))$ is a logistic sigmoid function and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ are the regression coefficients, that are computed by estimating the maximum likelihood of the logistic regression model on the scores of the LICIT set. The optimization is done using the *conjugate-gradient* algorithm [30].

Integrated PLDA system has the same structure as the baseline PLDA speaker verification system, but uses an extended training set. We assume the following hypothesis: under spoofing conditions, a PLDA model can better shape the intra-speaker and between-speakers variability if it is trained to discriminate, not only between multiple speakers like a baseline PLDA system does, but also between those speakers and all possible simulated versions of them. In this paper we study two types of voice coded speech: MCEP and LPC. We *always* use one type for the training and keep the other for testing to see how well the system is able to generalize. We apply the same approach to train a countermeasure (CM) system.

5. Experimental Results

Experimental Setup. The full i-vector extraction is done using *Spear* [31], an open-source speaker recognition toolbox based on *Bob* [32]. The UBM model is composed of 512 Gaussian components and is trained on NIST SRE04, SRE05, SRE06, Fisher and Switchboard datasets. For i-vectors, the rank of the total variability matrix is set to 400. For LDA, the projection matrix $\mathbf{A}$ is limited to 200 dimensions. For PLDA, the ranks of the subspaces $\mathbf{V}$ and $\mathbf{U}$ are set to 100 and 200, respectively. To train $\mathbf{T}$, $\mathbf{A}$, $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{W}$ (for WCCN) of the baseline system we use the data from SRE04, SRE05 and SRE06 (from which the test data used in our experiments were excluded).

For the matched and mismatched spoof conditions, the additional i-vectors are extracted from the synthetic MCEP or LPC coded versions of the training utterances using the same $\mathbf{T}$ matrix estimated for the baseline. Those i-vectors are then used to train the new whitening, LDA, WCCN and PLDA subspaces.

Results. On the LICIT protocol, the baseline PLDA system achieves an EER of 1.75% and a mDCF of 0.133 on pooled female and male trials, better than the baseline system presented in [10] (EER=2.99%, mDCF=0.154). On the SPOOF protocol, our baseline system obtains SFAR of 6.04% on the MCEP-coded speech trials (female + male), whereas an SFAR of 19.29% on the same trials is reported in [10]. The difference in performance between the two systems is possibly due to the additional use of LDA and WCCN, and also to a different implementation of acoustic feature extraction, voice activity detection, i-vector extraction and PLDA.

Table 3 reports the results on male trials of the cosine, PLDA and SVM scoring techniques on the anti-spoofing task, and for both matched (MCEP-MCEP, LPC-LPC) and mismatched (MCEP-LPC, LPC-MCEP) conditions. These results show no significant difference between the three methods. In the remainder of the experiments, we use the SVM classifier.

Fig. 2 reports the results on male and female trials of the SVM anti-spoofing classifier, and for both matched and mismatched conditions. It clearly shows that mismatch between training and testing conditions can lead to poor performance with accuracies bit higher than the chance level for LPC-MCEP condition (53.6% for male and 53.1% for female).

Table 2 reports the full results of the PLDA baseline, and both joint anti-spoofing and verification systems: score fusion, and integrated PLDA system. For the joint systems, the results when training with either MCEP and LPC coded speech are shown. To make the task realistic and more challenging, none of the spoofing trials was used to train score fusion. On the one hand, results indicate that score fusion can lead in some cases to good performance such as the training with LPC-coded speech (Row r3 in the table). However, this method does not generalize well: all the underlined SFAR values correspond to errors higher than the ones of the baseline. On the other hand, the integrated system is able to generalize fairly good. In all cases (Rows r4-r5 in the table), its SFARs are lower than the SFARs of the baseline. These results are confirmed in Fig. 3. This figure depicts the DET curves of female trials of both baseline and integrated system for which additional MCEP-coded speech are used during the training set. Further, both Table 2 and Fig. 3 show that error rates (EER and mDCF) of the integrated system are comparable and sometimes better than the baseline on LICIT protocol. This is not often the case in existing anti-spoofing approaches where the typical behavior is a trade-off between EER (on LICIT protocol) and SFAR [22].

Fig. 4 shows the score distribution of both baseline and integrated system on male trials. It indicates that the scores of the spoofing attacks are generally shifted to the left with the integrated system, leading to increased separation between genuine

Table 2: Performance summary on SRE06 speech conversion database. This table reports the EER (%) and minDCF on LICIT protocol, the SFAR (%) on MCEP voice converted trials, the SFAR (%) on LPC voice converted trials and the SFAR (%) of pooled MCEP and LPC voice converted trials.

		Female					***Male***					
System	**Additional training set**	**LICIT protocol**		**SPOOF protocol**			**LICIT protocol**		**SPOOF protocol**			
				MCEP	**LPC**	**pooled**			**MCEP**	**LPC**	**pooled**	
		EER	*mDCF*	*SFAR*	*SFAR*	*SFAR*	*EER*	*mDCF*	*SFAR*	*SFAR*	*SFAR*	
PLDA Baseline	-	1.76	0.133	6.13	10.84	8.48	1.60	0.166	9.19	15.28	12.24	r1
Score	MCEP	1.62	0.136	7.12	13.13	10.12	1.68	0.167	11.40	20.05	15.72	r2
Fusion	LPC	1.73	0.132	4.89	9.35	7.12	**1.49**	0.163	**2.83**	7.60	**5.21**	r3
Integrated	MCEP	**1.24**	**0.112**	**3.90**	5.82	4.86	1.78	**0.144**	4.42	6.01	**5.21**	r4
PLDA system	LPC	1.42	0.120	5.94	**2.97**	**4.46**	1.65	0.177	8.57	**2.03**	5.30	r5

and impostors trials.

Table 3: Comparison of anti-spoofing classification methods. This table shows the detection accuracy (%) on male trials for cosine, PLDA, and SVM methods.

	Cosine	PLDA	SVM
MCEP-MCEP	**92.2**	91.8	91.7
LPC-LPC	99.3	**99.4**	**99.4**
MCEP-LPC	98.3	**98.7**	98.3
LPC-MCEP	53.0	53.1	**53.6**

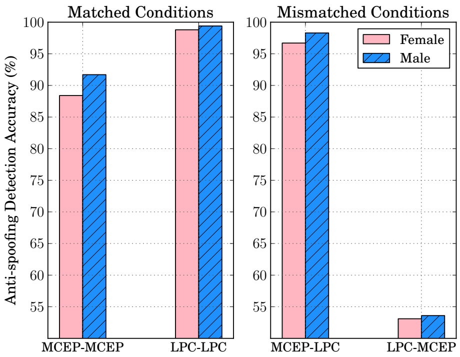

Figure 2: Spoofing detection accuracy on matched and mismatched conditions. This plot shows the accuracies (%) on spoofing trials for female and male speakers.

6. Conclusions

In this paper we introduce the i-vector paradigm to the task of spoofing detection. Further we present an integrated PLDA system for a joint operation of anti-spoofing and speaker verification. This system was found to generalize across two types of voice conversion attacks. Our experimental results suggest that the integrated system outperforms not only the baseline system, but also the score-fusion based approach, especially on the mismatched conditions between training and test.

Even if we addressed antispoofing in the face of mismatched vocoding techniques, namely MCEP and LPC, these are similar techniques originating from the same software package, SPTK. Thus, further experiments involving more severely mismatched spoofing techniques is required to claim truly generalized countermeasures. Nevertheless, we believe that our promising pilot experiments here are a possible candidate as a meaningful baseline system for voice anti-spoofing.

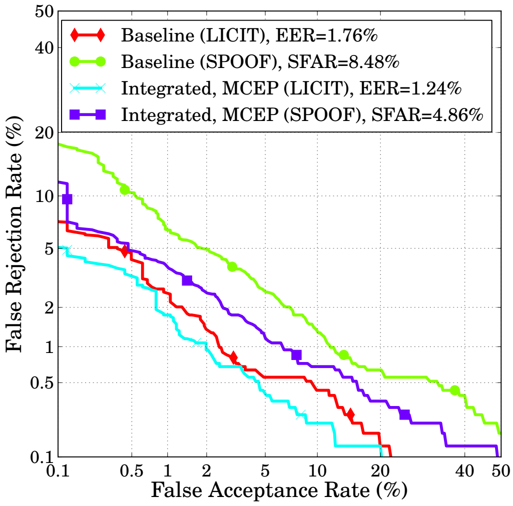

Figure 3: DET curves for female Trials. Results are for both Baseline system and Integrated system trained by MCEP-coded speech, and on both LICIT and SPOOF (pooled MCEP- and LPC-coded trials)

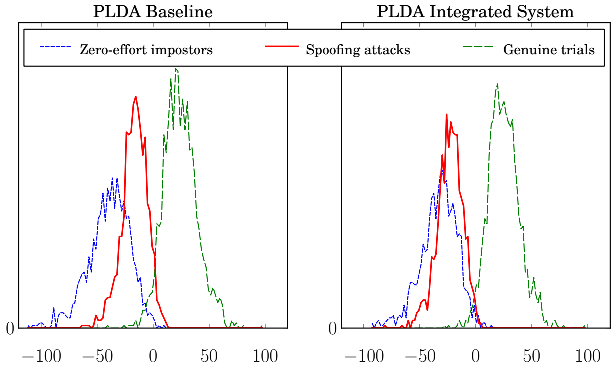

Figure 4: Score distribution for male trials. This figure shows the score distribution for both baseline and Integrated system trained on LPC-coded speech.

7. Acknowledgment

The research leading to this work has received funding from the Swiss National Science Foundation under the LOBI project, from Academy of Finland (proj. no. 253120 and 283256) and from the European Community's Seventh Framework Programme (FP7) under grant agreement 284989 (BEAT).

8. References

[1] A. Jain, A. Ross, and S. Pankati, “Biometrics: A tool for information security,” *IEEE Trans. on Information Forensics and Security (TIFS)*, vol. 1, no. 2, pp. 125–143, June 2006.

[2] N. K. Ratha, J. H. Connell, and R. M. Bolle, “Enhancing security and privacy in biometrics-based authentication systems,” *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

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

[4] Y. Stylianou, “Voice transformation: a survey,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[5] B. L. Pellom and J. H. Hansen, “An experimental study of speaker verification sensitivity to computer voice-altered imposters,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[6] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, “Voice forgery using ALISP: indexation in a client memory,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[7] D. Matrouf, J.-F. Bonastre, and C. Fredouille, “Effect of speech transformation on impostor acceptance,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

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

[9] T. Kinnunen, Z.-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. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[10] 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 *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[11] Z. Kons and H. Aronowitz, “Voice transformation-based spoofing of text-dependent speaker verification systems,” in *Proc. Interspeech*, 2013.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[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. European Signal Processing Conference (EUSIPCO)*, 2012.

[14] F. Alegre, R. Vipperla, N. Evans *et al.*, “Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals,” in *Proc. Interspeech*, 2012.

[15] Z. Wu, E. S. Chng, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” in *Proc. Interspeech*, 2012.

[16] D. L. P. L., M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, “Evaluation of speaker verification security and detection of HMM-based synthetic speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[17] F. Alegre, A. Amehraye, and N. Evans, “Spoofing countermeasures to protect automatic speaker verification from voice conversion,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[18] Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature,” in *ICASSP*, 2013, pp. 7234–7238.

[19] F. Alegre, A. Amehraye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[20] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE ICCV*, 2007, pp. 1–8.

[21] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[22] I. Chingovska, A. Anjos, and S. Marcel, “Anti-spoofing in action: joint operation with a verification system,” in *IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Biometrics*, 2013.

[23] D. A. Leeuwen and N. Brümmer, “Speaker classification I.” Springer-Verlag, 2007, ch. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pp. 330–353.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[25] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Interspeech*, 2011, pp. 249–252.

[26] L. E. Shafey, C. McCool, R. Wallace, and S. Marcel, “A scalable formulation of probabilistic linear discriminant analysis,” *IEEE Trans. in Pattern Analysis and Machine Intelligence*, 2013.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.

[28] S. Pigeon, P. Druyts, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1–3, pp. 237–248, 2000.

[29] N. Brümmer *et al.*, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Trans. on Speech, Audio and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[30] T. P. Minka, “Algorithms for maximum-likelihood logistic regression,” CMU Statistics Department, Tech. Rep. 758, 2001.

[31] E. Khoury, L. El Shafey, and S. Marcel, “Spear: An open source toolbox for speaker recognition based on Bob,” in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[32] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, “Bob: a free signal processing and machine learning toolbox for researchers,” in *ACM International Conference on Multimedia*, 2012.