



On the Need of Template Protection for Voice Authentication

Carlos Vaquero, Patricia Rodríguez

Agnitio S.L., Madrid, Spain

cvaquero@agnitio-corp.com, p.rodriguezgr@gmail.com

Abstract

In this work we study the need of template protection to provide security and privacy in text-dependent pass-phrase voice authentication systems. For this purpose, we analyze the robustness of two state-of-the-art speaker verification systems against attacks performed using input data generated from a compromised voice template. This analysis shows that compromised templates can be used to gain unauthorized access to authentication systems, when these systems use the same speaker verification technology, background models and pass-phrase as the one from which the compromised template was stolen. However we also show that the compromised template may not be helpful to attack an authentication system which uses a speaker verification technology or pass-phrase different from those considered in the system the template was obtained from. This fact facilitates the fulfillment of the main requirements that a protected template should meet to guarantee user privacy: irreversibility and unlinkability.

Finally we propose a set of guidelines for the design of voice authentication systems that enable the preservation of user privacy and provide revocability measures in case a template is compromised.

Index Terms: voice authentication, speaker recognition, template protection, security and privacy, revocability

1. Introduction

Biometric authentication is progressively replacing traditional authentication methods. The universality, uniqueness and permanence of biometric characteristics of individuals make biometric authentication natural, user friendly and difficult to deceive. However, these advantages pose a difficult challenge in security and privacy.

If a biometric template is compromised, revocability of the template can be limited or impossible, compared to traditional (knowledge or token based) authentication methods. As an example, in the case of fingerprint or iris recognition, revocability is limited since humans only have ten fingers and two eyes. Therefore, a compromised template is a much more serious security threat than a compromised password or secret cryptographic key. Moreover, since the compromised template represents an inherent and (almost) permanent biometric characteristic, it constitutes a threat for the user privacy as well. A typical example of privacy threat is cross-matching: a biometric template can be easily compromised in an authentication system with low security (e.g. gym access control) and used to access to a more critical application (e.g. bank accounts).

These security and privacy issues have motivated the research on template protection techniques in different fields of biometrics [1], [2], [3]. Recently, the ISO/IEC 24745:2011 standard [4] has been published to provide guidance for the protection of biometric information. This standard sets irreversibil-

ity and unlinkability as major requirements for protected biometric templates. Irreversibility refers to the impossibility to obtain the original biometric data from the protected template. Unlinkability refers to the possibility of generating different templates from the same biometric characteristic that cannot be used for cross-matching, enabling revocability of the biometric template.

Most of the previous research on template protection has been carried out for fingerprint [5] or iris biometrics [6], where revocability is very limited. Some techniques proposed in these fields include secure sketches [7], fuzzy commitment [8], [9], or cancelable biometrics [10], [11]. These techniques usually apply a non-invertible transformation K to the original biometric data, so that the irreversibility and unlinkability requirements are met at the cost of accuracy degradation due to the biometric information loss in the transformation. If the template is compromised, the user can be enrolled again and the template can be protected using a different transformation K' .

We could not find much published work on template protection for voice biometrics: in [12], random projections are proposed as cancelable biometric approach for speaker recognition, at the cost of certain accuracy loss. In [13], the authors extract binary templates that enable the use of fuzzy commitment schemes for template protection, originally designed for other forms of biometrics. They claim to obtain negligible accuracy loss while meeting the ISO/IEC 24745:2011 requirements. However it is arguable that binarization strategies would work correctly with state-of-the-art speaker recognition techniques.

In this work we analyze the need of template protection for text-dependent pass-phrase voice authentication systems. The motivation of this analysis is that, in our opinion, voice biometrics presents differences to other forms of biometrics. Particularly, we believe that, unlike iris or fingerprint biometrics, the speaker template obtained from a pass-phrase contains only a small amount of information of the physiological characteristic that it represents (i.e. vocal tract). This, combined to the fact that current speaker verification techniques use templates that hardly enable the reconstruction of the original user utterance, could be enough to ensure irreversibility and unlinkability (and thus revocability) in pass-phrase voice templates.

2. Experiment Protocol

In this study we analyze whether it is possible to use a stolen template to deceive a pass-phrase speaker verification system, in order to define the conditions under which template protection is needed in voice authentication systems. For this purpose, we consider two state-of-the-art speaker verification systems and we assume that a voice template from one of these systems has been compromised. The compromised template is then used to generate fraudulent input data to attack both systems.

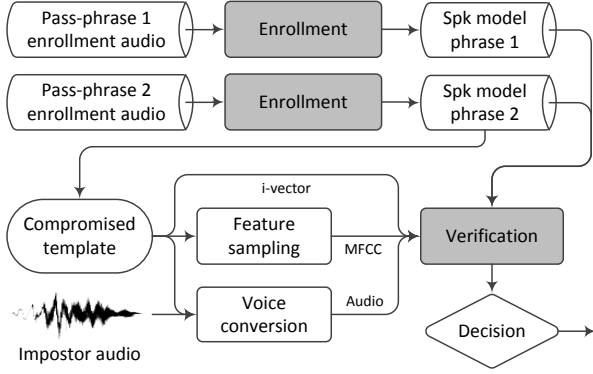


Figure 1: Diagram of attacks using a compromised template

2.1. Description of Speaker Verification Systems

We consider two approaches widely used in text-dependent speaker verification: a Hidden Markov Model (HMM) [14], [15] and an i-vector [16] with PLDA back-end [17], [18] systems. Both share the same Front End configuration, using 20 MFCC + Δ + $\Delta\Delta$ with feature warping [19]. The HMM system consist of a text-independent 128 component HMM-UBM that is adapted to the user enrollment data via MAP adaptation [20]. Thus, the speaker dependent HMM is adapted both to the speaker and the pass-phrase present in enrollment. The PLDA system uses a 256 component GMM-UBM and 300 total variability factors. The PLDA back-end uses a low rank Eigenvoice matrix (120 factors), and a full rank Eigenchannel matrix. Both systems are trained using text-independent telephone data from the NIST SRE 2004 to 2006 [21].

2.2. Description of attacks using a compromised template

We assume that a voice template for certain user and pass-phrase has been compromised, and we use it to generate input data for an authentication system, to gain unauthorized access. Figure 1 summarizes the attack process. We consider two different types of compromised voice templates:

- Speaker and phrase dependent i-vector: the voice template for the i-vector PLDA system is the i-vector obtained from the enrollment audios.
- Speaker and phrase dependent HMM: the voice template for the HMM system is the HMM adapted to the enrollment audios.

If one of these templates is stolen, it is possible to generate the following fraudulent data:

- The i-vector: If the compromised template is an i-vector, it can be used directly as input data to attack the PLDA system. The HMM system cannot be attacked this way.
- MFCC features: We can use the compromised template to sample MFCC features from the speaker model. The sampled features can be used to attack both systems.
- Audio: Using voice conversion techniques, it is possible to generate fraudulent input audio to attack both systems.

We use all generated data to attack both systems under analysis (when possible), so that we analyze whether cross-matching is feasible using templates from different technologies. Note also that the compromised template is text-dependent, and it may be used to attack authentication systems

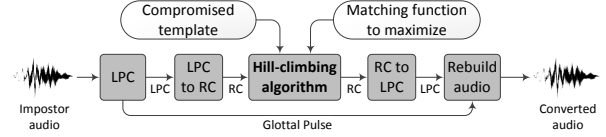


Figure 2: Diagram of the voice conversion process

that consider the same or a different phrase (depicted as phrases 1 and 2 in Figure 1).

2.3. Fraudulent data generation and injection

Below we describe the process to generate fraudulent input data from a compromised i-vector or HMM template. We also briefly comment the added difficulty of data injection, since some fraudulent data can be injected more easily than others.

2.3.1. i-vector generation and injection

The i-vector generation is straightforward and does not require any knowledge of the speaker verification system or its background models. However, its injection can be prevented securing the data transfer between the speaker verification modules of the voice authentication system, using encryption or integrity checking.

2.3.2. MFCC generation and injection

To generate features from an i-vector we obtain the speaker and phrase dependent GMM means from the UBM as follows [16]:

$$\mu_{SPK} = \mu_{UBM} + Tw \quad (1)$$

where μ_{UBM} is the UBM supervector, obtained concatenating all the UBM component means, μ_{SPK} is the speaker and phrase dependent GMM supervector, T is the total-variability matrix, and w is the compromised i-vector. We use μ_{SPK} and the UBM weights and covariances to build the speaker GMM, and we sample the fraudulent MFCCs from this GMM. In order to generate features from the HMM template we directly sample the MFCCs from the HMM. Note that, in general, this attack requires certain knowledge of the system and the background models. Even the HMM template is likely to contain the adapted HMM parameters only (i.e. component means and state transition probabilities), and thus the HMM-UBM parameters are required to sample the MFCCs properly.

The MFCC injection can also be prevented securing the data exchange between the different system modules.

2.3.3. Audio generation and injection

Fraudulent audio is generated using voice conversion techniques [22]. As shown in Figure 2, the proposed approach takes an impostor utterance, a stolen template and a matching function. First, LPC analysis is applied to the impostor utterance that contains the target phrase (the phrase required by the authentication system). The LPCs are converted to Reflection Coefficients (RCs) to ensure stability in the audio reconstruction. The RCs are then optimized frame by frame using a hill climbing approach based on the Uphill-Simplex algorithm [23]. The hill climbing approach modifies the RCs iteratively maximizing a matching function that takes the compromised template as input. The matching functions considered are:

- Scoring: The RCs are modified iteratively to maximize the verification score obtained by comparing the con-

Table 1: EER and normalized minDCF (x100) depending on the compromised template and the data generated from it

Compromised Template	Input data	attacking the PLDA system				attacking the HMM system			
		Same phrase		Different phrase		Same phrase		Different phrase	
		EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
baseline (no template)	impostor voice	8.98%	43.38	N/A	N/A	6.36%	29.51	N/A	N/A
i-vector	i-vector	50.00%	100.00	16.72%	65.77	N/A	N/A	N/A	N/A
	features	44.80%	100.00	5.69%	25.30	0.02%	0.03	0.01%	0.01
	audio max score	50.00%	100.00	11.50%	51.33	2.32%	4.04	1.88%	2.21
	audio min dist	49.35%	99.92	11.01%	45.67	6.24%	25.82	2.88%	5.72
HMM	features	1.37%	5.81	1.01%	4.31	48.13%	98.93	0.05%	0.14
	audio max score	7.25%	17.60	3.71%	11.87	49.91%	99.65	7.93%	21.15

verted audio and the stolen template. Note that complete knowledge of the speaker verification system is required to implement this matching function, that can be used both with i-vector and HMM templates as input.

- i-vector distance: The RCs are modified iteratively to minimize the distance between the stolen i-vector and the i-vector obtained from the converted audio. Thus, certain knowledge of the system is required to extract the i-vector from the converted audio. This matching function can only be used with i-vector template as input.

Audio injection can be achieved simply replaying the generated audio. Moreover, audio capture modules are sometimes not controlled by the voice authentication system, so software/hardware audio injection is harder to prevent.

2.4. Evaluation dataset description

In this study, we use the evaluation subset from the first part (pass-phrase scenario) of the RSR2015 database [24]. We do not use the development nor the background data, so that the systems are identical for all the phrases. The subset considered includes 30 phrases: 15 phrases are used for evaluation, while the 15 remaining phrases are used to simulate the case of a compromised template with a pass-phrase different from the one requested by the system under attack.

The scenarios proposed for this analysis are realistic and general in the sense that we do not consider matched background or development data, usually limited or unavailable, to train or adapt the speaker verification systems under test. Another reason to avoid the use of pass-phrase adapted systems in this study is that they may be harder to deceive using templates from different phrases, and we want to study the worst case scenario. On the other hand, avoiding the use of background and development data reduces significantly the accuracy of the systems under test, compared to state-of-the-art systems already evaluated on this database [15], [25], [26]. In any case, we believe that the baseline accuracy is not relevant for our purpose and does not affect the conclusions obtained in this work.

2.5. Evaluation measures

To evaluate the vulnerability of the systems under test against the fraudulent biometric data generated from compromised templates in different conditions, we run the RSR2015 evaluation replacing the impostor audios with the fraudulent data. This enables us to compute the EER or the minimum of the Detection Cost Function (minDCF) as defined in the NIST SRE 2008 [21], and compare directly to the baseline, which considers other speakers uttering the required pass-phrase as impostors. Therefore, a degradation in EER and minDCF with respect

to the baseline means that the compromised template is useful to attack the system: the fraudulent data generated is in general more competitive than the set of impostors in the database. On the other hand, an improvement in the accuracy means that the compromised template is not being helpful, since an impostor is in general more successful gaining unauthorized access than the fraudulent data generated. We do not consider cross-gender trials in this analysis.

3. Experimental results

Table 1 shows the EER and normalized minDCF (x100) obtained on the systems under test depending on the compromised template and the data generated from it. Results are presented for the cases of using a stolen template trained on the same phrase and on a different phrase from that required by the system to attack. We do not present results for the case of impostor speakers uttering a different phrase since it is not of interest in this study. Also, note that some attacks are not considered because they are not feasible (i.e. inject an i-vector into an HMM system).

Several interesting conclusions are extracted from the results in Table 1. First we observe that a compromised template that matches the phrase and the system under attack can be easily used to gain unauthorized access to the authentication system. The EER and normalized minDCF in these cases are close to 50% and 1 respectively, both for PLDA and HMM systems.

Another interesting conclusion is that speaker verification systems based on subspace methods for inter-speaker variability modeling may be more vulnerable to attacks using compromised templates of a different phrase than MAP/HMM based systems. Subspace and Factor Analysis [27] techniques enable adaptation on short utterances and the prediction of the speaker dependent feature distribution on regions of the space that are not seen during enrollment. Thus, they are probably more vulnerable to attacks using templates from different phrases than MAP based techniques. In addition, the information of the sequential structure of the data (phrase) that the HMM system captures is useful to prevent cross-phrase attacks.

Finally, note that none of the systems under analysis is vulnerable to attacks using a compromised template obtained from the other system, even if both the compromised and the attacked templates model the same phrase. This behavior is very interesting, since it shows that the information captured in the template strongly depends on the system, and this makes difficult to obtain useful biometric data from a compromised template.

The score mean and standard deviation depending on the nature of the input data for the PLDA and HMM systems are represented in Figures 3 and 4 respectively. These figures enable us to compare the score distributions in each case (under

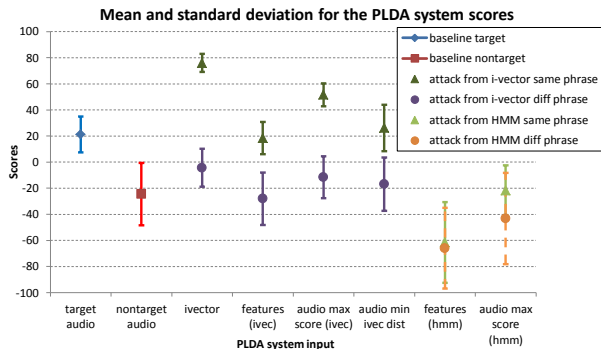


Figure 3: Score mean and standard deviation for the PLDA system, depending on the input data

Gaussian distribution assumption), confirming the conclusions previously enumerated. It is interesting to analyze the particular case of attacking an HMM system with audio generated from a HMM template with a different phrase. In Table 1, we observe that the EER for this case is worse than for the baseline, but the minDCF is not. In Figure 4 we can see that the scores obtained for this type of attack are in average less competitive than impostor scores, but the former show a relatively small standard deviation. This may cause a degradation in the accuracy of the system for high false alarm operating points, but an improvement for the low false alarm operating points, where the critical authentication applications usually operate.

4. Discussion: Security and Privacy

A compromised template represents a security and/or privacy threat for the user by itself, regardless of the data or service an unauthorized user can access using the template. The threat a compromised template involves depends on the scenario where the template is useful to deceive an authentication system. Table 2 shows the threat associated to each scenario.

Note that a compromised template constitutes a privacy threat only if it can be used to access other system and/or a system using other phrase. In fact, a voice template fulfills the irreversibility and unlinkability requirements if it can not be used to generate biometric data to attack a different system. Likewise, the template fulfills the unlinkability requirement if it cannot be used to attack a system with a different phrase.

Taking into account Table 2 and the results presented in Section 3, we can conclude that it would be possible to design a pass-phrase voice authentication system, without any template protection technique, following a set of guidelines:

- Using different speaker verification systems for different security levels or applications.
- Using different phrases for different security levels or applications. This implies also avoiding the use of phrases based on limited lexicon, and avoiding letting the user select his phrase, since the user may reuse the same or similar phrases in different applications.
- Providing change of phrase as revocability measure.
- Avoiding the use of subspace based speaker verification.
- Using software obfuscation techniques and protect/encrypt the background models, so that the attacker will not have enough knowledge of the system.

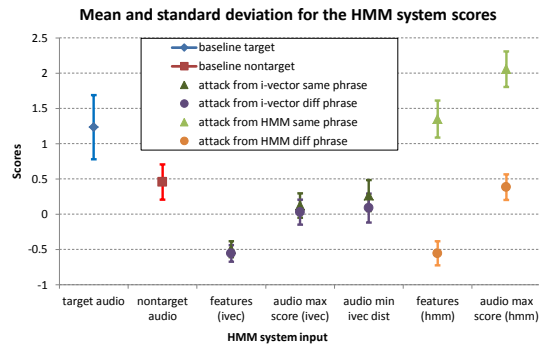


Figure 4: Score mean and standard deviation for the HMM system, depending on the input data

Table 2: Threat associated to each analyzed scenario

	Same phrase	Different phrase
Same system	Security only	Security and Privacy
Different system	Security and Privacy	Security and Privacy

- Protecting/encrypting/signing the data transmitted between system modules to avoid sniffing and injection.
- Using replay attack anti-spoofing systems [28], [29], so that the attacker cannot just replay the generated audio.

These guidelines mitigate the possible security and privacy threats that appear when a voice template is compromised. Regarding security, if these guidelines are followed, a compromised voice template would not be more critical than a compromised password or cryptographic key, that can be easily revoked. Regarding privacy, these guidelines are enough to guarantee irreversibility and unlinkability in the scenarios under study. A further analysis is required to determine up to which extent the definition of different systems or phrases apply.

5. Conclusions

In this work we have analyzed the conditions under which an attacker that makes use of a compromised voice template to generate input data for an authentication system would gain unauthorized access. We have considered two state-of-the-art speaker verification systems: PLDA and HMM, in order to obtain the compromised templates and also as authentication systems to attack. We have analyzed the robustness of these systems when different input data (audio, features or the template itself, when possible) is generated from the compromised template, considering same and cross-phrase situations and also the cases where the template is obtained from the same system to attack or from another system.

As general conclusion, we have observed that state-of-the-art pass-phrase voice authentication systems make use of templates that meet the irreversibility and unlinkability requirement up to a certain degree. A set of guidelines have been proposed so that the user privacy is preserved when his template is compromised, with no need of template protection. Additionally, these guidelines provide phrase modification as a revocability procedure.

Finally, note that this analysis focuses on pass-phrase authentication systems. Further research is needed to extend these conclusions to text-independent or other voice biometrics applications.

6. References

- [1] A. K. Jain, K. Nandakumar, and A. Nagar, "Biometric template security," *EURASIP J. Adv. Signal Process.*, vol. 2008, pp. 113:1–113:17, Jan. 2008.
- [2] C. Rathgeb and A. Uhl, "A survey on biometric cryptosystems and cancelable biometrics," *EURASIP J. Information Security*, vol. 2011, p. 3, 2011.
- [3] S. Rane, Y. Wang, S. C. Draper, and P. Ishwar, "Secure biometrics: Concepts, authentication architectures, and challenges," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 51–64, 2013.
- [4] ISO/IEC 24745:2011, "Information technology - security techniques - biometric information protection."
- [5] P. Tuyls, A. H. M. Akkermans, T. A. M. Kevenaar, G. J. Schrijen, A. M. Bazen, and R. N. J. Veldhuis, "Practical biometric authentication with template protection," in *AVBPA*, ser. Lecture Notes in Computer Science, vol. 3546. Springer, 2005, pp. 436–446.
- [6] S. Yang and I. Verbauwhede, "Secure iris verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, vol. 2, April 2007, pp. II–133–II–136.
- [7] Y. Sutcu, Q. Li, and N. D. Memon, "Protecting biometric templates with sketch: Theory and practice," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3-2, pp. 503–512, 2007.
- [8] A. Juels and M. Sudan, "A fuzzy vault scheme," in *Information Theory, IEEE International Symposium on*, 2002, pp. 408–.
- [9] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *SIAM J. Comput.*, vol. 38, no. 1, pp. 97–139, Mar. 2008.
- [10] A. Kong, K. H. Cheung, D. Zhang, M. S. Kamel, and J. You, "An analysis of biohashing and its variants," *Pattern Recognition*, vol. 39, no. 7, pp. 1359–1368, 2006.
- [11] A. B. J. Teoh, Y. W. Kuan, and S. Lee, "Cancelable biometrics and annotations on biohash," *Pattern Recognition*, vol. 41, no. 6, pp. 2034–2044, Jun. 2008.
- [12] A. B. J. Teoh and L.-Y. Chong, "Secure speech template protection in speaker verification system," *Speech Communication*, vol. 52, no. 2, pp. 150–163, Feb. 2010.
- [13] S. Billeb, C. Rathgeb, H. Reininger, K. Kasper, and C. Busch, "Biometric template protection for speaker recognition based on universal background models," *IET Journal on Biometrics*, 2015.
- [14] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [15] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, no. 0, pp. 56 – 77, 2014.
- [16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, August 2010.
- [17] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision (ICCV), IEEE 11th International Conference on*, Oct 2007, pp. 1–8.
- [18] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, June 2010.
- [19] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey Speaker and Language Recognition Workshop*, 2001.
- [20] J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.
- [21] NIST Speech Group, "Nist speaker recognition evaluation." [Online]. Available: <http://www.itl.nist.gov/iad/mig/tests/sre/>
- [22] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, vol. 1, May 2006, pp. I–I.
- [23] M. Gomez-Barrero, J. Gonzalez-Dominguez, J. Galbally, and J. Gonzalez-Rodriguez, "Security evaluation of i-vector based speaker verification systems against hill-climbing attacks," in *Interspeech*, August 2013, pp. 935–939.
- [24] A. Larcher, K. Lee, B. Ma, and H. Li, "RSR2015: database for text-dependent speaker verification using multiple pass-phrases," in *Interspeech*, 2012, pp. 1580–1583.
- [25] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and J. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," in *Interspeech*. ISCA, 2013, pp. 3684–3688.
- [26] A. Miguel, J. A. Villalba, A. Ortega, E. Lleida, and C. Vaquero, "Factor analysis with sampling methods for text dependent speaker recognition," in *Interspeech*, 2014, pp. 1342–1346.
- [27] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.
- [28] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.
- [29] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech*, 2013, pp. 925–929.