The Vicomtech spoofing-aware biometric system for the SASV challenge

Juan M. Martín-Doñas, Iván G. Torre, Aitor Álvarez and Joaquin Arellano

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)
{jmmartin, igonzalez, aalvarez, jarellano}@vicomtech.org

Abstract

This reports describes the proposed integration system for the spoofing-aware speaker verification (SASV) challenge. Our proposal mainly consists in the computation of a spoofing score using the speaker verification and spoofing embeddings (computed from dedicated networks) of the test utterance. This score is combined with the speaker verification score to obtain a spoofing-aware speaker verification score. The integration network is trained using a one-class loss function to focus on target trials. Our proposed system is evaluated in the ASVspoof19 database, showing competitive performance in comparison with other integration approaches.

Index Terms: speaker recognition, antispoofing, one-class learning

1. Introduction

Recent advances in deep neural network (DNN) architectures had led to significant improvements in the performance of speaker verification (SV) systems [1, 2, 3]. In parallel, similar advances have increasingly allowed the use of more sophisticated spoofing attacks that may include speech synthesis, replay attacks, or voice conversion. To deal with them, countermeasure (CM) systems have continuously integrated new approaches in increasingly complex spoofing attacks scenarios [3, 4]. Integrated SV-CM systems usually suffer significant performance degradation and are currently hot topic research in the context of biometric security systems [5, 6, 7, 8].

The Spoofing Aware Speaker Verification (SASV) Challenge 2022 aims to improve State of the Art (SOTA) robustness to both zero-effort impostor access attempts and spoofing attacks [9]. The SASV challenge focuses on evaluating the performance of integrated systems where both CM and SV subsystems are optimized together to improve the reliability of the full system in both scenarios. This closer to reality scenario is much more challenging and less explored than dealing with isolated cases of one type of attack.

The challenge baselines are based on the AASIST system [3], trained on ASVspoof2019 LA train partition [10], for CM, and ECAPA-TDNN [11] trained on VoxCeleb2 [12] dataset, for SV. AASSIST builds upon RawNet2-based encoder [13] to extract high-level representations from raw waveform inputs to feed a graph attention network [14] used for the extraction of CM embeddings. Meanwhile, ECAPA-TDNN is based upon Res2Net architecture [15] with a squeeze-excitation module to model channel interdependencies [16].

The goal of the challenge is to assess the performance of a joined system using an equal error metric (EER) without distinguishing between different speaker access or spoofed access attempts (SASV-EER). Besides, the system performance is also studied with more granularity considering a subset of target and non-target trials to estimate speaker verification performance (SV-EER) and another subset of spoofing attacks (SPF-EER). Only data from ASVspoof 2019 [10] and VoxCeleb 2 [12] dataset can be used for training and testing the systems, while results are performed on the ASVspoof 2019 evaluation partition. Two baseline systems are provided at the beginning of the challenge based upon the same pre-trained ASV and CM subsystems. *Baseline1* sums the scores produced by the subsystems, yielding poor performance. On the other hand, *Baseline2* is more elaborated, relaying on a three hidden layer fully connected neural network feed with three embeddings: two of them extracted from ECAPA on the enrolment and test utterances, while the third is extracted from AASIST test utterance.

In this work, we report an integrated system based upon the pre-trained AASIST and ECAPA-TDN models. First, a feed-forward vanilla neural network is trained to compute the cosine similarity between genuine speech and spoofing embeddings in order to obtain a spoofing score. Then, this score is linearly combined with the speaker verification score to obtain the final SASV score. With this approach, the SASV-EER is relatively reduced by 87% compared to *Baseline 2*, reporting a final SASV-EER of 0.84 %.

The remainder of this report is organized as follows. First, in Section 2, we detail the proposed spoofing-aware integration system. Then, in Section 3 the experimental framework and discussed results are presented. Finally, main highlights and conclusions are discussed in Section 4.

2. Proposed system

We propose an integrated SASV system based that combines an integration network followed by a combination of SV and SASV scores. First, the integration network uses the embeddings from the test utterance to compute a spoofing score, and then, this score is combined with the speaker verification (SV) score to obtain the final spoofing-aware speaker verification (SASV) metric.

Let us consider two different base systems trained for anti-spoofing and SV respectively, that compute a single embedding per utterance. Thus, we define $\mathbf{y}_{\text{sv}}$ as the SV embedding of the enrollment utterance, while $\mathbf{x}_{\text{sv}}$ and $\mathbf{x}_{\text{spf}}$ are the SV and spoofing embeddings from the test utterance, respectively. The SV score is directly obtained as the cosine similarity between the SV embeddings as $S_{\text{sv}} = \cos(\mathbf{y}_{\text{sv}}, \mathbf{x}_{\text{sv}}) \in [-1, 1]$.

Our objective is to compute a spoofing score for the test utterance that considers both embeddings from the base systems. Hence, we propose an integration network fed with the concatenation of both embeddings test: $\mathbf{x}_{\text{sv}}$ and $\mathbf{x}_{\text{spf}}$. This network is similar to *Baseline2* system of the SASV challenge, and it includes three feed-forward layers with LeakyReLU activations and 256, 128, and 64 hidden units, respectively. A batch normalization layer is added at the input of the deep neural network (DNN) to improve convergence during training by regularizing

the variance of the embedding units. Moreover, a linear layer is placed after the feed-forward layers to compute a new 64-dimension spoofing embedding $\mathbf{e}_{\text{spf}}$. Finally, a spoofing score is computed using the cosine similarity as $S_{\text{spf}} = \cos(\mathbf{w}, \mathbf{e}_{\text{spf}})$, where $\mathbf{w}$ is a vector network parameter representing the direction of genuine speech in the embedding space.

Finally, the SASV score can be obtained as a linear combination of the previously computed scores as follows,

$$S_{\text{sasv}} = \alpha S_{\text{sv}} + S_{\text{spf}}, \quad (1)$$

where α is a scalar value optimized during the network training phase. The integration network is trained to compute high S_{sasv} for target genuine trials. Inspired by previous works [17, 18], we use a one-class softmax loss function to focus on the target class. Given a batch of N trials, the loss is computed as

$$\mathcal{L}_{\text{OCS}} = \frac{1}{N} \sum_{n=1}^{N} \log \left(1 + e^{\beta \left(m_{z_n} - S_{\text{sasv,n}} \right) (-1)^{z_n}} \right), \quad (2)$$

where n is the batch trial index, $z = 0$ for the target class and one otherwise (non-target and spoof classes), β is a scale factor and m_z is a class-depending margin.

The idea behind our proposal is that a better spoofing score can be obtained when considering both the information contained in the SV and antispoofing test embeddings. Preliminary experiments considering other approaches –including only the spoofing embedding or the three different embeddings– yielded higher errors, supporting our previous hypotheses. While the use of the SV test embedding allows more robust decisions, the enrollment SV embedding does not yield improvements but degrades the performance of the integration system. On the other hand, the SV score is explicitly considered in the final score calculation. This allows the integration system to focus on especially difficult spoofing attacks (those that yield to high SV scores), while the SV information helps to discriminate them from zero-effort impostors and weaker attacks for the SV system. This strategy is also explored in [8], but instead of following a probabilistic framework we linearly combine the SV and spoofing scores, obtaining better results.

It is worthwhile to notice a relevant feature of our proposed system with respect to other integration approaches, as the one followed in *Baseline2*: the modularity of the SV system. Our system does not require the enrollment embedding to be processed by the integration network because the SV score is directly used in the final SASV score. Therefore, this system is compatible with different *homomorphic encryption* schemes [19], that allow certain operations as cosine similarity or linear score combination in the encrypted domain. Thus, the enrollment embeddings can be kept encrypted in the biometric system database, preventing unauthorized users to access private biometric information.

3. Experimental results

In this section, we describe the experimental framework and results obtained during the evaluation of our proposed system.

3.1. Experimental framework

Our proposed spoofing-aware speaker verification system is evaluated in the logistic access (LA) partition of the ASVspoof19 [10] database. The database is derived from the VCTK corpus [20], and it includes both bonafide speech and spoofing utterances generated by using different speech synthesis and voice conversion algorithms. The database is split into training, development, and evaluation subsets with non-overlapped speakers. Both the development and evaluation subsets include protocols for the evaluation of automatic speaker verification systems with spoofing attacks. Therefore, there are three different kinds of trials: target –bonafide test utterance from the same speaker as the enrollment one–, non-target –bonafide test utterance from an impostor speaker–, and spoof –synthetic test utterance–.

The base verification and antispoofing systems are the same as the ones used in the challenge baseline, that is, an ECAPA-TDNN for SV [11], trained using VoxCeleb2, and the AASIST network for antispoofing [3], trained using ASVspoof19.

The parameters selected for the loss function were $\beta = 20$. $m_0 = 0.9$ and $m_1 = 0.2$. The model was trained using the Adam optimizer [21] with a learning rate of $1e^{-4}$. A mini-batch of 24 trials was used. The architecture was trained during 20 epochs and the model with the best EER in the development set is kept for evaluation.

Our proposal is evaluated in terms of three different EER: SV-EER –target vs non-target trials–, SPF-EER –target vs spoof trials–, and SASV-EER –target vs non-target and spoof trials– . We compare our proposed system with the *Baseline2* integrated network and the probabilistic fusion framework proposed in [8]. Moreover, we also evaluated two additional integration approaches:

- A cascade antispoofing - SV approach that first detects spoofing utterances –spoofing score under a given threshold– and then computes the SV score. In this configuration, the threshold is selected by minimizing the error in the development set.
- A logistic regression approach that combines base systems SV and spoofing scores to compute a SASV score as the probability of being a target trial. The logistic regression is trained using the scores in the development set trials.

All the evaluated integration systems are based upon ECAPA-TDNN and AASIST respectively for SV and antispoofing tasks .

3.2. Results

It is now worth discussing the breadth and depth of the obtained results. Table 1 shows the EER obtained for the three evaluated tasks both for the the development and evaluation ASVspoof19 subsets. It is observed that our proposal outperforms the other methods, achieving the lowest EER in the three different evaluations tasks and in two tasks of the development development set. Overall, the results on the evaluation set are a 0.84% of SASV-EER, a 0.58% of SPF-EER and a 0.97% of SV-EER. The different spoofing-aware approaches perform better than the SV ECAPA baseline, but with different behavior. For example, the *Baseline2* obtains competitive results for spoofing detection –0.65% of SPF-EER– but degrades the SV performance –11.29% of SV-EER–. Both cascade and logistic regression approaches are optimized in the development set where they obtain good results –1.08% and 1.68% SASV-EER respectively–, but significantly degrade in the evaluation set –4.44% and 2.55% SASV-EER respectively–. This is especially true for the cascade system, where the spoofing threshold is not optimal for the evaluation set. The proposed system in [8] obtains competitive results all evaluation tasks –1.94%, 0.80% and 1.53% for

Table 1: *EER (%) results of our proposed integration system on ASVspoof 19, both development and evaluation sets. The results for the baseline systems and other approaches are also shown for comparison purposes.*

System	Development			Evaluation		
	SV-EER	SPF-EER	SASV-EER	SV-EER	SPF-EER	SASV-EER
ECAPA (SV) [11]	1.86	20.28	17.37	1.64	30.76	23.84
Cascade SPF-SV	1.89	0.42	1.08	1.64	6.59	4.44
Logistic regression	2.70	0.67	1.68	2.55	2.52	2.55
Baseline2 [9]	9.58	0.12	4.04	11.29	0.65	6.24
Zhang et al. [8]	2.02	**0.07**	1.10	1.94	0.80	1.53
Proposed system	**1.08**	0.13	**0.54**	**0.97**	**0.58**	**0.84**

SV, SPF and SASV respectively–. Nevertheless, the SV and antispoofing performance are still lower than the ECAPA model in SV-EER. Our proposed approach does not only perform better than previously published methods, but it also reduces the SV-EER of the ECAPA system while keeping spoofing detection performance similar to AASIST subsystem. This shows that our proposal can effectively exploit the information in the test embeddings along the SV scoring to obtain a competitive integrated spoofing-aware speaker verification system.

4. Conclusions

In this work, we have presented our proposed spoofing-aware speaker verification system for the SASV challenge. Our proposed integration network exploits the SV and spoofing embeddings of the test utterance to compute a robust spoofing score. The final SASV score is obtained as a linear combination of the SV score and the new spoofing score. The model is trained to compute higher scores for target trials using a one-class softmax loss function. Our proposed method shows competitive results in the ASVspoof19 development and evaluation sets, outperforming other spoofing-aware approaches, and performing significantly better than the base systems in the SV and spoofing detection tasks. As future work, we will test other base systems combined with our integration network, as well as different training strategies.

5. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 1962–1966.

[3] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *arXiv preprint arXiv:2110.01200*, 2021.

[4] J. M. Martín-Doñas and A. Álvarez, "The Vicomtech audio deepfake detection system based on Wav2Vec2 for the 2022 ADD Challenge," in *Accepted in 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2022.

[5] J. Li, M. Sun, X. Zhang, and Y. Wang, "Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss," *IEEE Access*, vol. 8, pp. 7907–7915, 2020.

[6] H.-j. Shim, J.-w. Jung, J.-h. Kim, and H.-j. Yu, "Integrated replay spoofing-aware text-independent speaker verification," *Applied Sciences*, vol. 10, no. 18, p. 6292, 2020.

[7] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[8] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," *arXiv preprint arXiv:2202.05253*, 2022.

[9] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV Challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

[10] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[12] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[15] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[18] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Log-likelihood-ratio cost function as objective loss for speaker verification systems," *Proc. Interspeech 2021*, pp. 2361–2365, 2021.

[19] A. Nautsch *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[20] J. Yamagishi. (2012) English multi-speaker corpus for CSTR voice cloning toolkit. [Online]. Available: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.