The Vicomtech Spoofing-Aware Biometric System for the SASV Challenge

Juan M. Martín-Doñais¹, Iván G. Torre²,³, Aitor Álvarez¹ and Joaquin Arellano¹

¹Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
²Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
³Department of Applied Mathematics, ETSIAE Universidad Politecnica de Madrid, Madrid, Spain

{jmmartin, aalvarez, jarellano}@vicomtech.org, ivan.gonzalez.torre@upm.es

Abstract
This paper describes our proposed system for the spoofing-aware speaker verification challenge (SASV Challenge 2022). The system follows an integrated approach that uses speaker verification and antispooﬁng embeddings extracted from specialized neural networks. Firstly, a shallow neural network, fed with the test utterance’s veriﬁcation and spoofing embeddings, is used to compute a spoof-based score. The ﬁnal scoring decision is then obtained by combining this score with the cosine similarity between speaker veriﬁcation embeddings. The integration network was trained using a one-class loss to discriminate between target and unauthorized trials. Our proposed system is evaluated over the ASVspoof19 database and shows competitive performance compared to other integration approaches. In addition, we compare our approach with further state-of-the-art speaker veriﬁcation and antispooﬁng systems based on self-supervised learning, yielding high-performance speech biometric systems comparable with the best challenge submissions.

Index Terms: speaker recognition, antispooﬁng, one-class learning, self-supervised, wav2vec2

1. Introduction
Recent advances in deep neural network (DNN) architectures have led to signiﬁcant improvements in the performance of speaker veriﬁcation (SV) systems [1, 2, 3], broadening its application in real scenarios such as access controls or assistant devices. However, these biometric systems can be vulnerable to spoofing attacks [4], which try to access the system by replicating the target speaker’s voice using different strategies such as voice synthesis and conversion, replay attacks, or even adversarial examples [5]. Therefore, countermeasure (CM) systems [6, 7] have been developed to try to detect and reject these attacks successfully. Robust deep learning-based approaches have been continuously integrated to deal with the increasingly complex spoofing attacks scenarios [8, 9, 10].

Despite the efforts on developing single systems for each task, the integration of complete biometric security systems (veriﬁcation and antispooﬁng) is still undergoing in the research community. Integrated systems usually suffer signiﬁcant performance degradation, especially when operating in tandem. Only a few works have explored the joint optimization of these systems, with strategies varying from the integration of models [11, 12, 13] to the use of joint architectures [14, 15, 16].

The Spoofing Aware Speaker Veriﬁcation (SASV) challenge 2022 [17] aims to improve state-of-the-art (SOTA) robustness to both zero-effort impostor access attempts and spoofing attacks. 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 complete system in both scenarios. Thus, the goal is to assess the performance of a joined system discriminating between different target speaker access and unauthorized accesses. This closer to reality scenario is much more challenging and less explored than dealing with isolated cases of one type of attack. Only data from ASVspoof 2019 [18] and VoxCeleb 2 [19] datasets can be used for training and testing the systems, while evaluations are performed on the ASVspoof 2019 dataset.

In this work, we present an integrated system based on pre-trained SV and antispooﬁng models. First, a feed-forward vanilla neural network was trained to compute a spoofing score from previously extracted SV and spoofing embeddings of the test utterance. Then, the ﬁnal SASV score is obtained from the SV and antispooﬁng ones using a linear operation. The integration model was trained using a one-class loss function focused on target trials, computing higher scores. We evaluated our approach using the challenge’s base SV and spoofing systems, reporting a 0.84% error rate for the SASV challenge. Moreover, we also evaluated more advanced SV and CM systems from the ones of the challenge, including self-supervised learning approaches based on wav2vec2 [20], yielding a competitive biometric system with a 0.14% error rate.

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

2. Proposed system
Our proposal for the development of a SASV system follows the integration approach. Thus, it relies on pre-trained speaker veriﬁcation and antispooﬁng neural networks for embedding extraction. Let us ﬁrst deﬁne ysv as the SV embedding of the enrollment utterance, while xsv and xsp are the SV and spoofing embeddings from the test utterance, respectively. The objective is to use three embeddings to compute the ﬁnal SASV score decision.

The proposed method is inspired by the Baseline2 of the challenge, which considers an integration neural network, fed with the three different embeddings, to discriminate target trials from non-target or spoofed ones. Instead, our integration network is trained to compute a spoofing score for the SV and antispooﬁng embeddings of the test utterance. The SASV score is then obtained as a linear combination between this spoofing score and the cosine similarity between the enrollment and test SV embeddings, allowing end-to-end system training.

Our proposal hypothesizes that a better performance can be obtained with the following two key aspects: the use of both SV and antispooﬁng test embeddings for the spoofing score and the...
explicit consideration of the SV score for the computation of the final decision. Although our pipeline considers a dedicated anti-spoofing system, the additional information provided by the SV embedding can help to learn other discriminative information from verification systems to detect spoofing artifacts. Preliminary experiments including only the spoofing embedding or the three different embeddings yielded higher errors, supporting our previous hypotheses. Furthermore, while the use of SV test embedding allows more robust decisions, the enrollment SV embedding degraded the performance of the integration system. Learning correlation information from the three different embeddings is a more challenging task for the computation of the SASV score than our proposed strategy. Finally, the SV score is explicitly considered in the final score computation. This strategy lets the integration system focus on challenging spoofing attacks (those that result in high SV scores). At the same time, the SV information helps to discriminate them from zero-effort impostors and weaker attacks for the SV system. This strategy is also explored in [13], but instead of following a probabilistic framework, we linearly combine the SV and spoofing scores, obtaining better performance.

The architecture of the integration network is similar to the Baseline2 system, and it includes three feed-forward layers with LeakyReLU activations and 256, 128, and 64 hidden units, respectively. This network is adapted as follows. First, the input to the network is formed with the concatenation of both test embeddings, $x_n$ and $x_{spf}$, followed by a batch normalization layer 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 $e_{spf}$. The output of the network is a spoofing score computed as the cosine similarity $S_{spf} = \cos(w, e_{spf})$, where $w$ is a trained vector network parameter representing the direction of genuine speech in the embedding space. Finally, the SASV score can be obtained as

$$S_{sasv} = \alpha S_{sv} + S_{spf},$$

where $S_{sv} = \cos(y_n, x_n)$ is the SV score and $\alpha$ is a scalar value optimized during the network training phase. The integration network is trained to obtain high $S_{sasv}$ for target genuine trials. Inspired by previous works [21, 22], we use a one-class softmax loss function. Given a batch of $N$ trials, the loss is computed as

$$L_{OCS} = \frac{1}{N} \sum_{n=1}^{N} \log \left(1 + e^{\beta (m_{zn} - S_{sasv,n})(-1)^{y_n}}\right),$$

where $n$ is the batch trial index, $z$ is 0 for the target class and 1 otherwise (non-target and spoof classes), $\beta$ is a scale factor, and $m_{zn}$ is a class-depending margin. Different margins allow broadening the expected scores for the attack classes (impostors and spoofing), thus increasing the robustness.

It is worthwhile to notice a relevant feature of our proposed system with respect to other integration approaches, such 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 [23], that allow certain operations as cosine similarity or linear score combination in the encrypted domain. Thus, the enrollment embeddings can be encrypted in the biometric system database, preventing unauthorized users from accessing private biometric information.

3. Experimental results

This section first describes the experimental framework, including the database used for training and evaluating our system, the trainable hyperparameters, and the evaluation metrics. Then, we reveal the results obtained for the SASV challenge 2022 and some additional evaluation scores using more discriminative SV and CM subsystems.

3.1. Experimental framework

Our proposed SASV system was evaluated in the logistic access (LA) partition of the ASVspoof2019 [18] database. The database is derived from the VCTK corpus [24], 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 development and evaluation subsets include protocols for evaluating 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 challenge baselines are on the AASIST system [9], trained on the LA train partition of the ASVspoof2019 for CM, and the ECAPA-TDNN model [2] trained on VoxCeleb 2 [19] dataset, for SV. AASIST integrates a RawNet2-based encoder [25] to extract high-level representations from raw waveform inputs in order to feed a graph attention network [26] used for the extraction of CM embeddings. Meanwhile, ECAPA-TDNN is based upon Res2Net architecture [27] with a squeeze-excitation module to model channel inter-dependencies [28].

The different approaches were evaluated in terms of three different equal error rate (EER) metrics, according to the kind of trials compared. These metrics are the SV-EER –target vs. non-target trials–, the SPF-EER –target vs. spoof trials–, and the SASV-EER –target vs. non-target and spoof trials–.

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 [29] with a learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, and a mini-batch of 24 trials. A scheduler is also considered to progressively reduce the learning rate at each iteration, as in Baseline2. The system was trained on the ASVspoof19 data for 40 epochs, keeping the model with the best SASV-EER in the development set for evaluation.

3.2. Challenge results

In this subsection, we compare our proposed system with the Baseline2 integrated network and the probabilistic fusion framework proposed in [13]. Moreover, we also evaluate two more integration approaches:

- A cascade CM-SV approach that first detects spoofing utterances –spoofing score under a given threshold-- and then computes the SV score. In this configuration, the threshold was 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 was trained using the scores from the development set trials.
Table 1: EER (%) results of our proposed integration system on ASVspoof19, 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) [2]          | 1.86        | 20.28      | 17.37     | 1.64     | 30.76     | 23.84     |
| Cascade CM-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 [17]          | 9.58        | 0.12       | 4.04      | 11.29    | 0.65      | 6.24      |
| Zhang et al. [13]       | 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**  |

Proposed system using ECAPA-TDNN and AASIST

(a) ECAPA-TDNN

(b) Proposed system using ECAPA and AASIST

Figure 1: Comparison between score empirical probability distribution functions for the ECAPA-TDNN verification system (top) and our proposed integration system (below). Performance scores are obtained by testing the evaluation set of ASVspoof19.

All the evaluated integration systems are based on ECAPA-TDNN and AASIST for SV and antispoofing tasks, respectively, using the allowed data according to the challenge rules.

Table 1 shows the EER obtained for the three evaluated tasks on the development and evaluation subsets of ASVspoof19. Our proposal outperforms the other systems, achieving the lowest EER in the three different evaluation plans and two tasks of the development set. Overall, the results on the evaluation set achieve a 0.84% of SASV-EER, a 0.58% of SPF-EER, and a 0.97% of SV-EER. Thus, our system achieves 7th position among 23 submissions to the challenge1. The different spoofing-aware approaches perform better than the SV ECAPA baseline but with different behaviors. The Baseline2 obtains competitive results for spoofing –0.65% of SPF-EER– but degrades the SV performance –11.29% of SV-EER–. Both cascade and logistic regression approaches were optimized in the development set, obtaining good results –1.08% and 1.68% SASV-EER, respectively–, but significantly degraded in the evaluation set –4.44% and 2.55% SASV-EER respectively–. This is particularly significant for the cascade system, where the spoofing threshold may not be optimal on the evaluation set. The proposed approach in [13] obtains competitive results on all evaluation tasks –1.94%, 0.80% and 1.53% for SV, SPF, and SASV, respectively–. Nevertheless, the SV and antispoofing performance are still lower than the ECAPA model in SV-EER. Our proposed system outperforms previously published methods and reduces the SV-EER of the ECAPA system while keeping spoofing detection performance similar to Baseline2 approach. These results demonstrate that our proposal can effectively exploit the information in the test embeddings along the SV scoring to obtain a competitive SASV system.

In addition, Figure 1 shows the score empirical probability distributions for the ECAPA-TDNN (top) compared with our proposed approach (below). As it can be observed, there is a high overlap between the target and spoof score distributions in ECAPA, revealing the lack of robustness against spoofing attacks. On the other hand, our proposed system achieves better separation between target scores and the remaining classes, translating to lower EERs, as can be seen in Table 1.

3.3. Evaluation using SOTA SV and CM subsystems

In order to further evaluate the benefits of our proposed system, in this subsection, we evaluated our approach using different SOTA base SV and CM subsystems. For speaker verification, two additional networks were considered:

- **ECAPA2**: The ECAPA-TDNN previously evaluated but, in this case, trained using speech data from both Voxceleb 1 and 2 datasets.
- **W2V2-SV**: This method is based on the self-supervised learning paradigm, using a wav2vec2 (W2V2) architecture [20]. We evaluated the pipeline system proposed in [31], which combines W2V2 with ECAPA-TDNN.

1https://sasv-challenge.github.io/challenge results/
Table 2: EER (%) results for evaluating our proposed integration system using different SV and CM subsystems. Results for the ECAPA2 and W2V2 base SV subsystems, and the best SASV challenge submission, are also included for comparison purposes.

| Subsystem Development | Evaluation |
|-----------------------|------------|
|                      | SV-EER | SPF-EER | SASV-EER | SV-EER | SPF-EER | SASV-EER |
| ECAPA2               | 1.21   | 17.86   | 15.23    | 0.76   | 27.01   | 20.63    |
| W2V2-SV              | 0.81   | 14.80   | 12.39    | 0.50   | 20.42   | 15.73    |
| ECAPA AASIST         | 1.08   | 0.13    | 0.54     | 0.97   | 0.58    | 0.84     |
| ECAPA2 AASIST        | 0.88   | 0.20    | 0.46     | 0.45   | 0.58    | 0.52     |
| ECAPA2 W2V2-CM       | 0.68   | 0.00    | 0.33     | 0.49   | 0.26    | 0.35     |
| W2V2-SV AASIST       | 0.73   | 0.07    | 0.40     | 0.45   | 0.46    | 0.45     |
| W2V2-SV W2V2-CM      | **0.40** | **0.00** | **0.13** | **0.18** | **0.11** | **0.14** |
| Best SASV system [30] | 0.00   | 0.00    | 0.00     | 0.11   | 0.17    | 0.14     |

Figure 2: Score empirical probability distribution functions of the integrated system using wav2vec2 architectures for both SV and CM subsystems. The performance is computed by testing the evaluation set of ASVspoof19.

downstream model, with the models finetuned along using Voxceleb 2 data.

Regarding the antispoofing task, we included our previously proposed approach based on W2V2 (W2V2-CM) [10]. This model comprises a pre-trained wav2vec2 feature extractor and a spoofing network classifier trained on ASVspoof19. We chose the best model for our evaluations in ASVspoof 2021 LA [32], which used the cross-lingual W2V2 model pre-trained with 128 languages and narrowband FIR filtering data augmentation during model finetuning. The spoofing embedding was selected from the attentive pooling layer instead of the final linear layer, as these embeddings showed better robustness in our preliminary experiments. Moreover, the SV and spoofing embeddings of the evaluated subsystems were first length-normalized before feeding them to the network, thus avoiding norm scale issues.

Table 2 shows the different EER obtained when evaluating several combinations of these base subsystems with our proposed integration approach. The results for the ECAPA2 and W2V2-SV systems, without using spoofing countermeasures, as well as the best SASV challenge submission [30] are also shown for comparison purposes. The two considered SV subsystems perform better than the ECAPA baseline in SV and antispoofing subtasks. Moreover, the integration approach with these systems and W2V2-CM further increases the performance. In particular, the W2V2-based subsystems outperform the remaining ones, with the combination of both W2V2 models yielding a SASV-EER of 0.14% in the evaluation set. This is a very competitive result comparable with the best approach presented in the SASV challenge. The winner system is based on robust SV and CM integrated systems, scoring normalization, and an additional ensemble classifier. On the other hand, we achieve similar results with a more straightforward integration approach by relying on high-performance W2V2 subsystems.

Finally, Figure 2 shows the empirical probability distribution function of scores obtained for our proposed integrated approach along with the W2V2-based subsystems. This system gets a significantly better separation between the three different classes, with the target trials achieving the highest scores and the spoofing trials the lowest scores.

4. Conclusions

In this work, we have presented our proposed spoofing-aware speaker verification system for the SASV challenge. Our integration network exploited the SV and spoofing embeddings of the test utterance to compute a robust spoofing score. The final SASV score was obtained as a linear combination of the SV and spoofing scores. Moreover, the model was trained to compute higher scores for target trials using a one-class softmax loss function. This method showed satisfactory 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. Furthermore, the performance increased when combining our integration approach with robust W2V2-based SV and CM subsystems, yielding competitive results among the top systems in the SASV challenge. As future work, we will evaluate the finetuning of the complete SASV system and different training strategies.

5. Acknowledgements

This work was supported in part by the Spanish Centre for the Development of Industrial Technology (CDTI) through the Project EGIADA—RED DE EXCELENCIA EN TECNOLOGÍAS DE SEGURIDAD Y PRIVACIDAD under Grant agreement CER20191012.
6. 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), 2018, pp. 5329–5333.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in Proc. Interspeech 2020, 2020, pp. 3830–3834.

[3] N. J. M. S. Mary, S. Unesh, and S. V. Katta, “S-Vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoders,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 404–413, 2022.

[4] C. B. Tan et al., “A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction,” Multimedia Tools and Applications, vol. 80, no. 21, pp. 32 725–32 762, 2021.

[5] 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), 2018, pp. 1962–1966.

[6] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” Speech Communication, vol. 66, pp. 150–153, 2015.

[7] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, “ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech,” IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 252–265, 2021.

[8] J. Monteiro, J. Alam, and T. H. Falk, “Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers,” Computer Speech & Language, vol. 63, p. 101096, 2020.

[9] 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,” in 2022 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), 2022, pp. 6367–6371.

[10] J. M. Martín-Doñuous and A. Álvarez, “The Vicomtech audio deepfake detection system based on Wav2vec2 for the 2022 ADD Challenge,” in 2022 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), 2022, pp. 9241–9245.

[11] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, “Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion,” Proc. Interspeech 2018, pp. 77–81, 2018.

[12] 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.

[13] Y. Zhang, G. Zhu, and Z. Duan, “A probabilistic fusion framework for spoofing aware speaker verification,” in Proc. Odyssey 2022, 2022, pp. 77–84.

[14] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, “Joint speaker verification and antispooﬁng in the t-vector space,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 821–832, 2015.

[15] 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.

[16] 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.

[17] 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,” in Proc. Interspeech (submitted), 2022.

[18] 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, pp. 101114, 2020.

[19] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in Proc. Interspeech 2018, 2018, pp. 1086–1090.

[20] A. Baevski, A. Zhou, H. and Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12 449–12 460, 2020.

[21] 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.

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

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

[24] 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

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

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” in International Conference on Learning Representations, 2017.

[27] 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.

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in International Conference on Learning Representations, 2015.

[30] A. Alenin, N. Torgashov, A. Ohkhotnikov, R. Makarov, and I. Yakovlev, “A subnetwork approach for spoofing aware speaker verification,” in Proc. Interspeech (submitted), 2022.

[31] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, “Large-scale self-supervised speech representation learning for automatic speaker verification,” in 2022 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), 2022, pp. 6147–6151.

[32] J. Yamagishi et al., “ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection,” in Proc. 2021 ASVspoof Workshop, 2021, pp. 47–54.