A Comparative Study of Fusion Methods for SASV Challenge 2022

Petr Grinberg[1], *Vladislav Shikhov*[1]

[1]Samsung R&D Institute Russia

`p.grinberg@samsung.com, v.shikhov@samsung.com`

Abstract

Automatic Speaker Verification (ASV) system is a type of biometric authentication. It can be attacked by intruder, who falsifies data in order to get access to protected information. Countermeasures (CM) are special algorithms that detect this spoofing-attacks. While ASVspoof Challenge series were focused on development of CM for fixed ASV system, new Spoofing Aware Speaker Verification (SASV) Challenge organizers believe that best results can be achieved if CM and ASV systems are optimized jointly [1].

One of approaches for cooperative optimization is a fusion over embeddings or scores obtained from ASV and CM models. Baselines of SASV Challenge 2022 present 2 types of fusion: score-sum and back-end ensemble with 3-layer MLP. This paper describes our research of other fusion methods, including boosting over embeddings, which has not been used in anti-spoofing studies before.

Index Terms: SASV Challenge 2022, Countermeasure, Automatic Speaker Verification, Fusion

1. Introduction

Automatic Speaker Verification (ASV) system is a biometric authentication method working with humans' speech. As any type of protection it may be susceptible to tampering, called spoofing-attacks, which can be classified into 4 kinds: impersonation, voice conversion (VC), replay attacks and text-to-speech (TTS) [2]. There are two possibilities, how to deal with spoofing-attacks. First of them is improve reliability of ASV systems. Second – use specialized algorithms, called countermeasures (CM), which detect unauthorized access.

ASVspoof initiative and first ASVspoof Challenge 2015 were created in order to standardize research in the field, metrics and databases [3]. This competition was held 4 times in 2015, 2017, 2019 and 2021 respectively. Last time it contained 3 tasks [4]:

- Logical Access (LA) task: spoofing data is created by VC and TTS algorithms and transmitted over telephone and VoIP networks.
- Physical Access (PA) task: CMs for replay attacks.
- Speech Deepfake (DF) task: combination of other tasks, but without speaker verification.

New Spoofing Aware Speaker Verification (SASV) Challenge 2022 is a continuation of ASVspoof series. However, in ASVspoof CMs work with fixed ASV system. Organizers of SASV Challenge believe that join optimization of CM and ASV systems can lead to more robust models. This paper compares different approaches for solution for this challenge.

Existing architectures for CM systems can be divided into two types: ones, which work with raw signal as input, and ones, which apply time-frequency transforms and create spectrograms. In [5] authors showed that spoofing artifacts lie in different frequency sub-bands rather than in full-band. Because of that, performance of spectrogram-based methods relies on used time-frequency algorithm's resolution in sub-band, where spoofing attack left marks. Different spectrograms enhance different frequencies, what leads to their contrasting capabilities. Hence, combining architectures with different front-ends can result in robust model. Raw-input solutions define which frequencies are helpful by themselves during train stage. However, their aptitude is still limited. Thus, architecture with great discriminate power should use fusion over model with different input types.

Fusion methods appeared in literature are usually done over scores obtained from different CM models. It was shown that non-linear types of fusion perform better than the linear ones [6]. However, average or weighted sum with pre-normalization are still the most common [1, 7, 8, 9, 10]. Other appeared methods, where scores from CMs are stacked in one feature-vector, are presented below:

- Support Vector Machines (SVM) with different type of kernels: linear, residual basis function (RBF) and polynomial. It was used in [6, 11, 12].
- Logistic Regression was tested by authors of [6, 12, 13]
- Gaussian Mixture Model (GMM) appeared in [6].
- MLP, Decision Tree and Random Forest were used as a final classifier over future-vectors in [12].

All of this methods, except Decision Tree, are shown to improve the performance and surpass single model systems.

Fusion over embeddings occurs much less often. Authors of [14] created Fusion-Layer, which combines outputs of model sub-parts. SASV Challenge 2022 organizers used 3-layer MLP with embeddings from both ASV and CM systems [1].

In this paper we explore and compare performance of different fusion methods. Apart from existing solutions, which mostly use scores, we take embeddings from CM and ASV systems. Moreover, only one ensemble method of Decision Trees was applied in anti-spoofing research – bagging (specifically, Random Forest). We propose usage of another Decision Tree ensemble method called boosting. Concretely, fusion is done using CatBoost [15], and it significantly outperforms SASV Challenge 2022 baselines on all metrics. In addition, we test usage of Random Fourier Features (RFF) [16] with logistic regression base, which is a RBF SVM approximation method.

The remainder of this paper is organized as follows. Section 2 describes CM and ASV systems, which embedding will be taken for fusion. Section 3 describes experimental setup, which results are described in Section 4. Conclusions are presented in Section 5.

2. CM and ASV systems

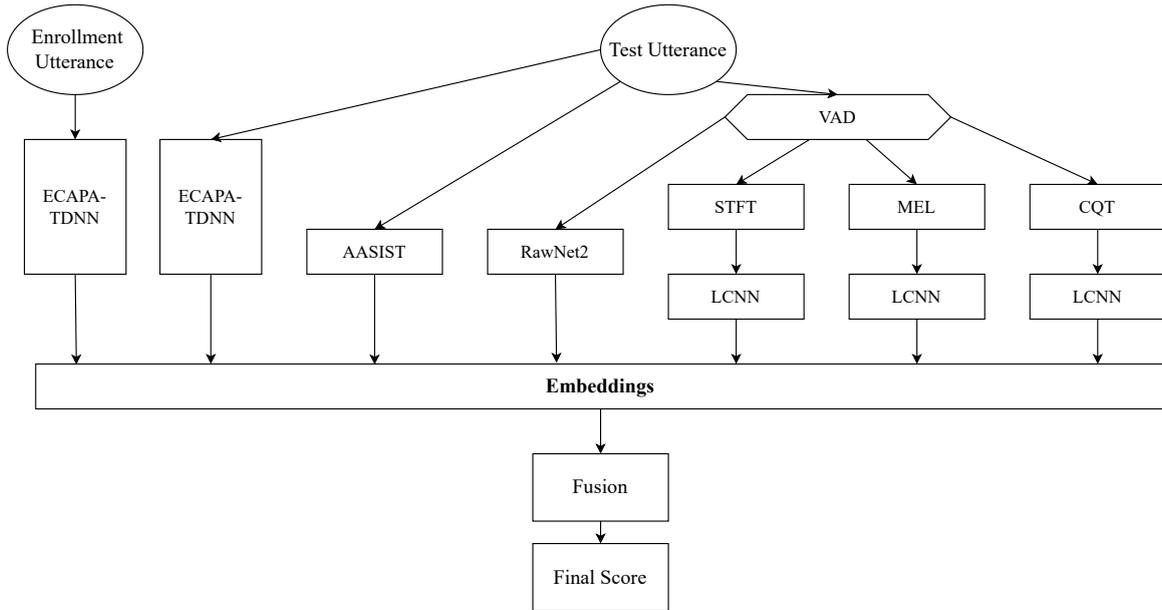This section describes what CM ans ASV systems were used in experiments.

Figure 1: *Pipeline for fusion over embeddings experiment*

2.1. Countermeasure

For CM system we took 5 different models: two with raw input and three with spectrogram front-ends.

Firstly, we used AASIST [17], which is a baseline CM subsystem in SASV Challenge 2022. This is spectro-temporal graph attention network working with raw input. Sending waveform through AASIST results in 190-dimensional embedding. Scores are outputs of last Fully-Connected layer (using softmax is an option).

Secondly, we adopted RawNet2 [11] architecture to our training pipeline. Specifically, Dropout layers were added. RawNet2 was a baseline CM system for ASVspoof Challenge 2021 [4]. It consists of Sinc-layer, ResBlocks and GRU-layer, ending with two FC-layers. Embeddings are 1024-dimensional vectors obtained from output of first classifying FC-layer. Scores comes out the last layer.

Thirdly, we took LightCNN (LCNN) architecture proposed in [18] with 3 types of input as was done in [19]: STFT, MEL and CQT spectrograms. This time-frequency transform methods enhance different frequency sub-bands, which allows to create a robust model by combining them. LCNN is a classic CNN model with Max-Feature-Map operation as activation functions. Embeddings are outputs of penultimate FC-layer and have 80 dimensions for each type. Scores comes out the last layer.

2.2. Automatic Speaker Verification

For ASV system we used ECAPA-TDNN proposed in [20], which is also a baseline ASV subsystem in SASV Challenge 2022. This subsystem consist of Squeeze-Excitation Res2Blocks, Statistic Pooling and Multi-layer Feature Aggregation. ECAPA-TDNN returns two 160-dimensional vectors for enrollment and test utterances respectively. Score is a cosine similarity between embeddings.

3. Experimental Setup

In this section we describe how we trained our models and used fusion over them to following evaluation.

3.1. Training Stage

Our ASV system ECAPA-TDNN was pre-trained on Vox-Celeb2 dataset [21]. It is same to the SASV Challenge 2022's organizers ASV subsystem.

All CMs were trained on ASVspoof 2019 LA train partition [22]. We also took pre-trained AASIST as was done in competition. However, RawNet2 and LCNN-based models were trained from scratch and had data preprocessing with data augmentation.

In DF task of ASVspoof challenge speech waveforms were encoded and then decoded with different lossy codecs. This added some distortion into data and complicated detection of spoofing-attacks. However, in practice compressed audio is common. So we believe that robust CM, which can be used in real life situations, should handle this encoding-decoding perturbations. For this we added random compression with mp3 and aac codecs as data augmentation.

There is a recent study, which shows that ASVspoof Challenge Database has very uneven distribution of silence: the bonafide samples has much longer silences, than many of attacks, especially TTS ones [23]. This leads to a serious problem: models partially make their predictions based on duration of silence in data. It was shown that same architectures learned on samples with trimmed silence get much worse results. We believe that strong biometric authentication system has to deal with no silence data and reject samples with hush: indeed, if somebody wants access to protected information he or she should provide his or her biometry, and there is no biometry in silence. Therefore, we preprocess data for RawNet2 and LCNN-based models by removing silences using simple magnitude-based Voice Activity Detection (VAD) algorithm to improve their detecting spoofing artifacts capabilities.

We trained RawNet2 and LCNN-based models for 30

Table 1: *Results of Fusion Methods over embedding on Dev and Eval sets. Best value in column is bold. SASV Challenge 2022 Baselines' results and ECAPA-TDNN without any CM are at the bottom of the table.*

Fusion Method	Metric					
	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
3-layer MLP	48.58	49.08	**0.06**	0.85	15.88	25.10
Logistic Regression	50.13	50.03	0.17	1.50	16.59	25.41
SVM (Linear Kernel)	49.88	49.33	9.13	16.07	17.70	27.67
SVM (RBF Kernel)	27.02	22.99	0.26	1.32	11.00	12.24
SVM (Polynomial Kernel)	29.14	31.76	7.41	10.70	13.92	19.39
RFF (Logistic Regression Based)	37.53	36.24	2.76	8.73	14.11	21.08
GMM	50.56	49.95	49.90	49.47	50.04	49.64
Random Forest	35.42	32.73	0.06	0.76	13.13	18.16
CatBoost	3.62	3.87	**0.06**	0.61	**2.09**	**2.90**
Score-Sum (Baseline 1)	32.88	35.32	**0.06**	0.67	13.07	19.31
3-layer MLP (Baseline 2)	12.87	11.48	0.13	0.78	4.85	6.37
ECAPA-TDNN	**1.88**	**1.63**	20.30	30.75	17.38	23.83

epochs with Adam optimizer.

3.2. Fusion methods over embeddings

We already discussed why fusion of CMs' scores with different inputs is better than the single system. We hypothesize that at the same manner fusion of CMs with and without silence trimming results in well-performing model, which detects artifacts and do not make predictions based only on silence duration.

For this study we get embeddings from each sub-model and concatenate them into one long feature-vector, which is used as input for final classifier. Then classifier is trained on embeddings from train data. RawNet2 and LCNN-based models' embeddings are taken from preprocessed and encoded-decoded data. For each LCNN-based model input spectrogram is split into 3 equal-sized parts. We get embeddings for each part and stack them together.

Final scores for development and evaluation sets are received from trained classifier. We again used same preprocessing and encoding-decoding for RawNet2 and LCNN-based models.

Resulting pipeline of our system is given in Figure 1. Final feature-vector is a 2288-dimensional vector of numerical features.

Methods tested as final classifier are as follows:

- 3-layer MLP. Same architecture, that was used in SASV Challenge 2022 Baseline, but with our 2288-dimensional vector as input: 3 FC-layers with 256, 128 and 64 out features respectively, LeakyReLU with 0.3 negative slope as activation and one final classifier FC-layer.

- Logistic Regression. Iterations are limited to 1000 (we also tried bigger value, but it's performance was worse). Regular coefficient is $\lambda = \frac{1}{m}$, where $m = 25380$ is the amount of utterances in ASVspoof 2019 LA train partition. All other parameters are set to default[1].

- SVM with Linear kernel. Iterations are limited to 50000, regular coefficient is $\lambda = \frac{1}{m}$. All other parameters are set to default[1].

- SVM with RBF kernel. Iterations are limited to 50000, regular coefficient is $\lambda = \frac{1}{m}$. All other parameters are set to default[1].

- SVM with Polynomial kernel. Iterations are limited to 50000, regular coefficient is $\lambda = \frac{1}{m}$, degrees are set to 7. All other parameters are set to default[1].

- RFF with logistic regression. Classifier uses linear principal component analysis (PCA) with 1024 dimensions and Standard Scaler. Number of random features is set to 5000. Iterations of inner logistic regression are limited to 50000, regular coefficient is $\lambda = \frac{1}{m}$. All other parameters are set to default[1].

- GMM. Number of components is set to 2. Amount of EM iterations is 1000. All other parameters are set to default[1].

- Random Forest. Number of estimators is 1000, all other parameters are set to default[1].

- CatBoost. Number of estimators is 700, all other parameters are set to default (We used catboost v.1.0.4 [15]).

3.3. Fusion methods over scores

For completeness of our study we tested if we could improve the results by doing fusion over scores. First of all, we replicated SASV Challenge 2022 Baseline 2 (it will be called RB2 for shortness). Then we took RB2 and our best model (OBM for shortness) from embeddings experiment and stacked their scores for each utterance in one 4-dimensional feature vector.

Then we tested same classifiers, except MLP and RFF, because of too small input size. For Logistic Regression and SVM-based methods iterations were unlimited and regular coefficient was set to 10000.

4. Results

The results for different fusion methods over embeddings described in Section 3.2 are shown in terms of SV-EER, SPF-EER and SASV-EER in Table 1. Performance of SASV Challenge 2022 Baseline systems and ECAPA-TDNN without any CM

---
[1]We used sklearn v.0.24.2 [24]

Table 2: *Results of Fusion Methods over scores on Dev and Eval sets. Best value in column is bold. RB2, OBM, SASV Challenge 2022 Baselines' results and ECAPA-TDNN without any CM are at the bottom of the table.*

Fusion Method	Metric					
	SV-EER		SPF-EER		SASV-EER	
	Dev	Eval	Dev	Eval	Dev	Eval
Logistic Regression	3.81	3.57	**0.06**	**0.52**	2.20	**2.75**
SVM (Linear Kernel)	3.81	3.57	**0.06**	**0.52**	2.20	**2.75**
SVM (RBF Kernel)	3.90	3.69	**0.06**	**0.52**	2.15	2.79
SVM (Polynomial Kernel)	3.79	3.54	**0.06**	0.53	2.14	2.77
GMM	8.19	6.75	2.49	2.19	5.53	5.29
Random Forest	7.42	5.37	1.19	0.87	3.63	3.39
CatBoost	6.73	4.87	0.19	0.68	2.83	2.88
RB2	13.38	11.97	**0.06**	0.68	6.12	6.25
OBM	3.62	3.87	**0.06**	0.61	**2.09**	2.90
Score-Sum (Baseline 1)	32.88	35.32	**0.06**	0.67	13.07	19.31
3-layer MLP (Baseline 2)	12.87	11.48	0.13	0.78	4.85	6.37
ECAPA-TDNN	**1.88**	**1.63**	20.30	30.75	17.38	23.83

are at the bottom of the table. SV-EER measures how model distinguish target and not-target trials. SPF-EER – target and spoofed trials. SASV-EER is a general metric, where non-target and spoofed trials are treated equally.

We can clearly see that GMM approach works as a random binary classifier. It was expected, because embeddings are hardly similar to Gaussian vectors. Hence, GMM can not be used for fusion methods over embeddings.

Unlike the results from [6] where Polynomial SVM outperformed RBF and Linear ones, in our experiment SVM with RBF Kernel shows best of three results. Logistic regression has similar results for RBF SVM and Poly SVM in terms of SPF-EER and SASV-EER respectively, but SV-EER is equal to random classification. Linear SVM's quality indicators are poor too. Reason for such performance of Logistic Regression and Linear SVM is a linear inseparability of bonafide and non-target trials, which can be observed by superiority of non-linear approaches. 3-layer MLP has the same problem, caused by small train set and narrow layers with large input size.

Random Fourier Features should approximate SVM with RBF kernel in theory. We can see from the Table 1 that RFF indeed has similar performance to SVM methods, but it has much worse results, than RBF SVM.

Random Forest, which is bagging ensemble method of Decision Trees, exceeds the results of Baseline 2 in terms of SPF-EER and has similar results as Polynomial SVM in other metrics. Thus, Random Forest is better than SVM with polynomial kernel in our study.

Final tested method is CatBoost. It surpasses all competitors methods in all metrics with a huge margin in SV-EER and SASV-EER. Moreover, it reduces error-rates, obtained by baselines, by a relative $71.8\%, 66.28\%, 0\%, 8.95\%, 56.90\%, 54.47\%$ for SV-EER, SPF-EER and SASV-EER on Development and Evaluation sets respectively. Only single ECAPA-TDNN outperforms CatBoost approach on SV-EER. It is not surprising because using countermeasures worsen verification ability.

To further improve the results we took model with CatBoost fusion method (OBM) and combined it with RB2 in our second experiment described in Section 3.3. It's results are presented in Table 2.

Scores are more likely to be from Gaussian Mixture than embeddings. GMM approach still deteriorate performance of best single model, however it is far from random classification. Hence, GMM is not hopeless as score fusion method.

Logistic Regression and SVM-based methods have alike results. Polynomial SVM has the best of four performance with only slightly impairment in terms of SPF-EER and SASV-EER on Evaluation Set. This four methods outperform OBM on evaluation set and slightly worsen on development set. Hence, this methods work as some kind of regularization.

Both Decision Tree ensemble methods got unsatisfactory quality indicators, however, they are better than GMM.

Thus, our best results are obtained using SVM with 7 degree polynomial kernel score fusion of RB2 and OBM, which is a fusion of ECAPA-TDNN, AASIST, RawNet2 and three LCNN models' embeddings using CatBoost. This system is our submission for SASV Challenge 2022.

5. Conclusions

This papers reports a comparison of different fusion methods over embeddings. Results show that CatBoost approach, which did not appear in anti-spoofing studies before, outperforms all other methods by a huge margin. Moreover, we trimmed silence in audio data for some of our sub-models, what was demonstrated to worse performance in [23], but still have desirable results. This indicates how robust is the CatBoost method. Other fusion over embeddings approaches can still get satisfactory results, but only for distinction of target and spoofing-trials. They have poor results on SV-EER and SASV-EER metrics.

However, for fusion over scores everything is different. Logistic Regression and SVM-based methods outperform best single system on Evaluation set and act like a regularization. Random Forest and CatBoost did not get great results. Thus, this methods are better to use with embeddings than scores. Finally, GMM should be used only as score fusion approach, and it's performance highly depend on distribution of scores from sub-models. In addition, our study confirms that non-linear fusion methods are better than the linear ones.

6. References

[1] 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.

[2] 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. 130–153, 2015.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[4] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[5] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 333–340.

[6] ——, "Spoofing Attack Detection Using the Non-Linear Fusion of Sub-Band Classifiers," in *Proc. Interspeech 2020*, 2020, pp. 1106–1110.

[7] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection," in *Proc. Interspeech 2021*, 2021, pp. 4259–4263.

[8] J. Cáceres, R. Font, T. Grau, J. Molina, and B. V. SL, "The Biometric Vox system for the ASVspoof 2021 challenge," in *Proc. ASVspoof2021 Workshop*, 2021.

[9] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop Labs' Submission to the ASVspoof 2021 Challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[10] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion," in *Proc. Interspeech 2017*, 2017, pp. 17–21.

[11] 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.

[12] W. H. Kang, J. Alam, and A. Fathan, "CRIM's system description for the ASVspoof 2021 Challenge," in *Proc. ASVspoof 2021 Workshop*, 2021.

[13] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "UR channel-robust synthetic speech detection system for ASVspoof 2021," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 75–82.

[14] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[16] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, ser. NIPS'07. Curran Associates Inc., 2007, p. 1177–1184.

[17] 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.

[18] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[19] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 1033–1037.

[20] 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.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[22] 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.

[23] N. M. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?" in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 55–60.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.