# Deepfake audio detection by speaker verification

Alessandro Pianese, Davide Cozzolino, Giovanni Poggi and Luisa Verdoliva
University Federico II of Naples, Italy
Email: {name.surname}@unina.it

*Abstract*—Thanks to recent advances in deep leaning, sophisticated generation tools exist, nowadays, that produce extremely realistic synthetic speech. However, malicious uses of such tools are possible and likely, posing a serious threat to our society. Hence, synthetic voice detection has become a pressing research topic, and a large variety of detection methods have been recently proposed. Unfortunately, they hardly generalize to synthetic audios generated by tools never seen in the training phase, which makes they unfit to face real-world scenarios. In this work we aim at overcoming this issue by proposing a new detection approach that leverages only the biometric characteristics of the speaker, with no reference to specific attacks. Since the detector is trained only on real data, generalization is automatically ensured. The proposed approach can be implemented based on off-the-shelf speaker verification tools and we test several such solutions on three popular test sets, obtaining good performance, high generalization ability and high robustness to audio impairment.

*Index Terms*—Deepfake audio, synthetic voice, spoofing detection.

## I. INTRODUCTION

Recent methods for synthetic voice generation are becoming more and more sophisticated. State-of-the-art methods for Text to Speech (TTS) synthesis and Voice Conversion (VC) allow one to generate extremely realistic audio waveforms with such a high level of naturalness to fool both humans and automatic speaker verification (ASV) systems [1]. Convincing fake audios, when used for malicious purposes, pose a serious threat to our society, as they allow us to impersonate people or create deepfake videos simply by synchronizing the generated speech with the movement of the lips.

In the current literature, several detection methods have been proposed, most of them based on supervised learning. Traditional methods typically rely on the extraction of acoustic features in the spectral domain. In fact, through time-frequency analysis it is possible to obtain a rich and expressive representation of the voice, based on the well-known frequency cepstral coefficients or constant Q cepstral coefficients [2], [3]. Also effective are first- and second-order spectral features [4] as well as higher-order polyspectral features [5], exploiting the observation that synthesized speech can introduce uncommon spectral correlations. Other approaches use features related to auto-regressive modeling of the speech [6] or a combination of spatial and spectral features [7], [8], or rely on a constrained convolutional neural network where the first layer act as a bank of band-pass filters with a fixed rectangular-shaped filter response and learned cut-in and cut-off frequencies [9].

These approaches are usually very effective on the data they are trained for, but suffer a significant, even dramatic, performance impairment when used on data synthesized by generation models never seen in training. This is a major weakness in real-world scenarios, where the manipulation attack is not known in advance. Therefore, it is extremely important to devise methods that keep working well also "in the wild". Indeed, one major goal of the 2019 ASVspoof initiative was to foster spoofing countermeasures that generalize smoothly to unseen attacks. To this end, a dataset was created where the training and development partitions included a set of 6 attacks, while the evaluation partition included a different set of 13 attacks only partially overlapping with the former.

Given the fast progress of text-to-speech technologies, it is important to understand to what extent methods trained for specific attacks generalize to new ones, and try to ensure a good behavior for them also in real-world scenarios. A possible way to improve generalization is to consider a large number of attacks in training or, equivalently, to fuse multiple detection systems, each tuned to a different attack. However, this amount only to moving the boundary a little further, without really solving the problem. The system will keep being prone to new unseen attacks.

A more radical solution to the generalization issue is to rely on one-class methods [10]. By definition, these are trained only on the class of interest, the real class in our case, and data that depart from it are considered as anomalies, that is, fake. Since no fake data are used in training, there is no dependence on any specific form of manipulation, and generalization is ensured by default. Of course, a description of the class of real audios so precise to enable reliable anomaly detection is out of reach. However, the idea of training only on real data can be implemented in other forms. For example, several papers have recently appeared [11]–[14] where video deepfakes are detected by means of a person-of-interest approach. The detector tries to establish if the subject of the video is really the person that is claimed to be: if not, the video is labeled as fake. Therefore, rather than looking for traces of manipulation, these methods leverage high-level biometric features. Since training takes place only on original data, generalization is guaranteed. In [11]–[13] only video biometric features are exploited, while [14] exploits both audio and video features.

Here, we investigate the potential of the person-of-interest (PoI) approach when only the audio track is available or exploitable. We assume to know the identity of the speaker of the audio under analysis and to have a reference set of pristine audio tracks for such identity. Then, the identity is verified by comparing suitably extracted embedded vectors of test and reference audios. We note that this differs from [15], where the proposed supervised approach is fine-tuned on
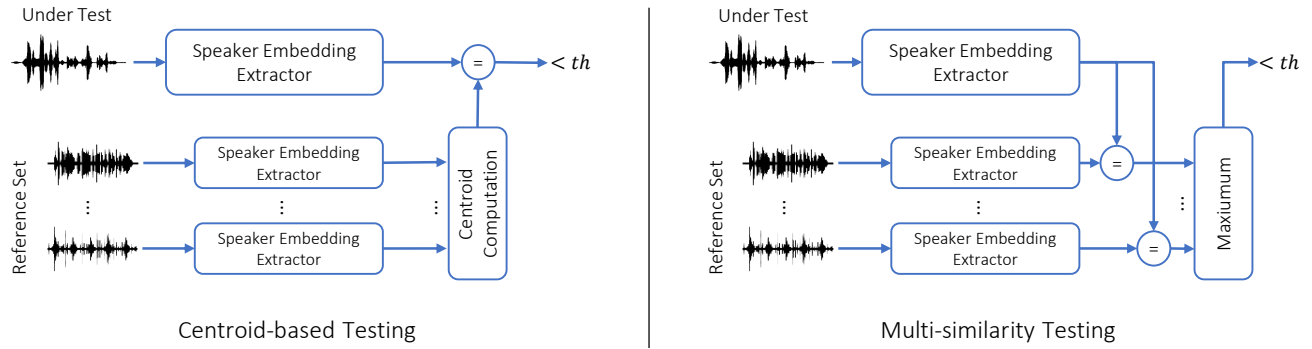
Fig. 1. Schemes of the two testing strategies: Centroid-Based (CB) and Multi-similarity (MS). In both strategies, we project the audio under-test and all the audios of the reference-set in the embedding space by the speaker embedding extractor. On the left we show the CB testing strategy: a centroid vector is computed averaging all the embedded vectors of the reference-set and the embedded vector of the audio under-test is compared only with the centroid one. The similarity obtained in this way is used as decision statistic to detect deepfake audios. On the right we show the MS testing strategy: the embedded vector of the audio under-test is compared with each embedded vector of the reference-set and the maximum similarity is provided as decision statistic.

real and synthetic speech of the identity under test. Instead, it can be observed that, with our approach, synthetic audio detection boils down to a speaker verification problem. This is true, and in fact we adapt several state-of-the-art speaker verification methods to perform fake audio detection and study their performance in several conditions of interest.

In summary, our main contributions are:

- we show that person-of-interest methods, including speaker verification approaches, can be used to effectively detect synthesized speech, even in the presence of noise;
- we prove this solution to largely outperform state-of-the-art supervised methods in terms of generalization ability.

## II. SYNTHETIC AUDIO DETECTION WORKFLOW

We want to evaluate the potential of audio speaker verification techniques for deepfake audio detection. In our scenario, the *claimed* identity of the speaker is known, and a reference set ($\mathcal{R}$) of real audios of the same identity is available. Our procedure is to compare the audio under test with those of the reference set in the embedding space of the audio speaker verification. To this aim, we consider two different testing strategies, shown in figure 1: Centroid-Based (CB) and Multi-Similarity (MS). The first strategy is inspired to Centroid-Based Metric Learning [19]. We compute the centroid $c = \sum_{i \in \mathcal{R}} r_i / |\mathcal{R}|$ of the reference set embedding vectors $\{r_i\}_{i \in \mathcal{R}}$. Then, the similarity, $s = d(x, c)$, between this centroid and the embedding vector of the audio under test, $x$, is used as decision statistic, and the audio is labeled as fake if $s$ is below a given threshold. In all cases, $d(\cdot, \cdot)$ is the similarity metric adopted by the speaker verification technique. In the Multi-Similarity testing strategy, proposed in [13], [14], a similarity index, $s_i = d(x, r_i)$, is evaluated between the audio under test and each audio of reference set. Then, only the maximum $s = \max_{i \in \mathcal{R}} s_i$ is considered as a decision statistic.

We consider various audio speaker verification techniques, they are described below and a summary of their characteristics is reported in table I.

The **Clova-AI** [16] method is named after the proposing group, involved in the development of the Clova personal assistant for Android. The main focus is on metric learning, considered especially well-fit to deal with a realistic open-set scenario. In input, the model accepts a 40-dimensional Mel Spectrogram with a 25ms window width and a 10ms frame shift. The backbone is Thin ResNet-34 which is the same as ResNet-34 [20] except for using only one-quarter of the channels in each residual block. Further minor modifications lead to Fast ResNet-34, with reduced complexity. Self-Attentive Pooling (SAP) [21] is used to aggregate frame-level features into utterance-level representation, with larger weights associated to more informative frames. The authors introduce also a new loss function, called Angular Prototypical (AP) loss, that chooses some elements of the batch to obtain centroids for each speaker, and uses cosine distance to obtain a similarity score for each sample.

In [17], researchers of the same group report on their proposals for the 2020 VoxCeleb Speaker Recognition Challenge (VoxSRC). The best experimental results are obtained with the **H/ASP** method with AP+Softmax loss. 64-dimensional log Mel-filterbanks are now used as the input to the network. The backbone is again a variant of ResNet-34 with one-half (as opposed to one-quarter) channels in each residual block. Attentive Statistic Pooling (ASP) [22] is used, where the channel-wise weighted standard deviation concurs with the weighted mean for temporal-frame aggregation. The AP loss proposed in [16] is combined with a standard softmax loss. The method uses two types of augmentation: additive background noise and reverberation simulation. In the first case, background audios (ambient sounds, human speech, and music) from the MUSAN dataset [23] are used, with a random signal-to-noise-ratio (SNR) going from 5 to 20dB. For reverberation simulation, the simulated room impulse response filters provided in [24] are used.

The **ECAPA-TDNN** [18] method was submitted as a solution for the 2020 VoxSRC Workshop. It is based on a Time Delay Neural Network (TDNN), with several improvements with respect to previous solutions. First, Squeeze-and-Excitation (SE) blocks are introduced to explicitly model

| Method | Input | Architecture | Pooling | Aug. | Loss function | Feature len. | Similarity |
|---|---|---|---|---|---|---|---|
| [16] ClovaAI | 40-MelSpec | Thin ResNet-34 | SAP | No | AP | 512 | Cosine Similarity |
| [17] H/ASP | 64-MelSpec | Half ResNet-34 | ASP | Yes | AP+softmax | 512 | Cosine Similarity |
| [18] ECAPA-TDNN | 80-MelSpec | SqueezeExcitation ResNet | ASP | Yes | AAM-softmax | 192 | Cosine Similarity |
| [14] POI-Forensics | 257-Spec | GN ResNet-50 | GAP | No | Contr. loss | 256 | Squared Euclidean |
| [14] POI-Forensics + aug | 257-Spec | GN ResNet-50 | GAP | Yes | Contr. loss | 256 | Squared Euclidean |

TABLE I
CHARACTERISTICS OF USED SPEAKER VERIFICATION TECHNIQUES.

channel interdependencies. Together with convolutional, block normalization and ReLU layers, these are included in the core SE-Res2Blocks residual blocks. Then, ASP is employed to select the most insightful features, and the Angular Margin Softmax (AAM-softmax) [25] is used as loss function. As input, the model accepts 80-dimensional Mel Spectrogram, again with a 25ms window width and a 10ms frame shift. Augmentation strategies similar to those of H/ASP are used.

**POI-Forensics** was proposed in [14] to perform deepfake detection based on audio-visual information with a person-of-interest (PoI) approach. A deepfake is detected when the person in the video turns out not to be who is claimed to be. In other words, deepfake detection is explicitly recast as a person identification problem, based on soft biometric information, with decision relying jointly on video and audio clues. Here, we neglect the video information and use exclusively the audio track, performing therefore speaker verification. In [14], the audio feature extraction network is based on a ResNet-50 backbone with Group-Normalization and Global Average Pooling. A crucial role is played by the contrastive learning formulation, aimed at embedding vectors of the test audio close to other embedded vectors of the same subject, but far from those of different subjects. In particular, a Distance-Based Logistic Loss is adopted, proposed in [26] and widely used in the literature. As input, the network accepts a 257-dimensional Spectrogram with a 25ms window width and a 10ms frame shift. Here, we consider also a variant of the method with augmentation, obtained by adding background noise with random SNR going from 3 to 25dB.

## III. EXPERIMENTAL RESULTS

In this Section, we describe the supervised methods for audio spoofing detection used as state-of-the-art references, the datasets used in the experiments, and the experimental results, with special emphasis on generalization and robustness.

### A. Reference methods

**RawNet2-antispoofing** [9] processes the waveform by a bank of band-pass filters. Afterwards, this model makes use of residual blocks but with two key differences: firstly, each residual block contains a filter-wise feature map scaling to enhance the discrimination ability of the features, as shown in [27]; secondly, before the fully connected layer, it is used a gated recurrent unit to aggregate features to form a single utterance-level representation.

**RawGAT-ST-mul** [7] is a model that aims to exploit artefacts in specific subbands and temporal portions of a recording. Also this model uses a bank of band-pass filters, whose output is first fed to several residual blocks and then is given as input to the separable spectral-temporal attention graph layer. The resulting features are merged by element-wise multiplication before being fed to the unified spectral-temporal attention layer.

**LCNN-LSTM-sum-p2s** [2] is based on the light convolutional neural network proposed in [28]. It has been improved by adding two bidirectional LSTM layer with skip connections. This method accepts as input 60-dimensional Linear Frequency Cepstral Coefficients (LFCCs) with a frame window of 20ms and a frame shift of 10ms.

### B. Datasets

To evaluate the methods described above, we selected three public datasets that provide information about the subject.

**ASVSpoof2019** [29] was released as part of the 2019 Automatic Speaker Verification Spoofing And Countermeasures Challenge. The dataset is divided into different parts, and we consider only the evaluation set formed by audio tracks manipulated with 19 different speech synthesis and voice conversion methods. The selected part contains 7,355 pristine and 63,882 manipulated utterances from 67 different subjects. To build the reference sets, we adopt a leave-one-out strategy, that is, all the real utterances of the POI are included in the reference set except the audio under test.

**FakeAVCelebV2** [30] is an audio and video deepfake dataset created from the videos of the VoxCeleb2 dataset [31]. Videos are divided into four self-explanatory categories: RealVideo-RealAudio, FakeVideo-RealAudio, RealVideo-FakeAudio, and FakeVideo-FakeAudio. Fake speech tracks are generated using the real-time voice cloning tool SV2TTS [32]. We extracted audio tracks only from the RealVideo-RealAudio and RealVideo-FakeAudio categories, since the same audio tracks are associated with fake videos. Each category contains 500 files, so we totalled 1000 audio clips. For the reference set, we use 10 pristine audios of the subject of interest drawn from VoxCeleb2.

The **In-The-Wild Audio Deepfake dataset (IWA)** [33] contains 20.8 hours of pristine recordings and 17.2 hours of fake audios, for a total of 31,779 files, belonging to 54 English-speaking identities. Fake audios were not generated by the authors but collected from publicly available material.

To ensure consistency, the authors collected real material of the same type as the fake one, e.g., a fake speech from Richard Nixon prompted them to gather a real speech from the same politician. Overall, the dataset contains 19,963 pristine utterances and 11,816 manipulated ones of 54 different subjects. Also for this dataset we adopt the leave-one-out strategy to build the reference sets.

### C. Experimental analysis

In Table II, we report the results on the three datasets in terms of Equal Error Rate (EER), Area Under the ROC Curve (AUC), and minimum normalized tandem Detection Cost Function (t-DCF) [34]. The t-DCF is the weighted sum of the missing detection rate and the false alarm rate. We use the weights provided in the ASVspoof 2019 challenge [35] and report the minimum t-DCT value over the decision threshold. Note that all the speaker verification techniques are trained on the VoxCeleb2 dataset [31], while the supervised methods are trained on the ASVSpoof2019 [29] training dataset. For all the techniques, we analyzed 4 seconds of the audio signal.

On the ASVSpoof2019 dataset, the supervised baselines obtain a near perfect performance, which is not surprising, as they were developed for this dataset. However, the picture changes completely when other, off-training, datasets are considered. Only LCNN-LSTM-sum-p2 keeps providing an acceptable performance, while the other two methods degrade to the coin-tossing level. On the contrary, the proposed framework proves to work almost equally well across all the datasets, with a generally good and sometimes excellent performance, which degrades somewhat only on the IWA dataset, due to its intrinsically more challenging nature. In this latter case, the Multi-similarity testing strategy seems to be preferable, while the Centroid-Based testing strategy ensures a better performance on "simpler" datasets. This can be explained by noting that audios collected in the wild (from social networks or video streaming platforms) are likely to have a more scattered distribution in the embedding space and evaluating the similarity for each reference audio may increase the probability of a good matching. POI-Forensics provides the best overall performance, but the speaker verification methods are not far, especially H/ASP, which is a remarkable result, considering that they were adapted on the fly to perform deepfake audio detection.

In figure 2, only for the IWA dataset, we show the AUC performance as a function of the number of elements of the reference set. As expected, the performance increases with increasing cardinality. However, with Centroid-Based testing strategy the performance plateaus at about 20 elements, while it keeps growing steadily with the Multi-Similarity testing strategy, confirming the above observations.

### D. Robustness analysis

In figure 3, we show the AUC performance on the IWA dataset in the presence of background noise. We consider real-world noise using the environmental sounds of the ESC-50 dataset [36]. In detail, we consider the sounds: breathing,
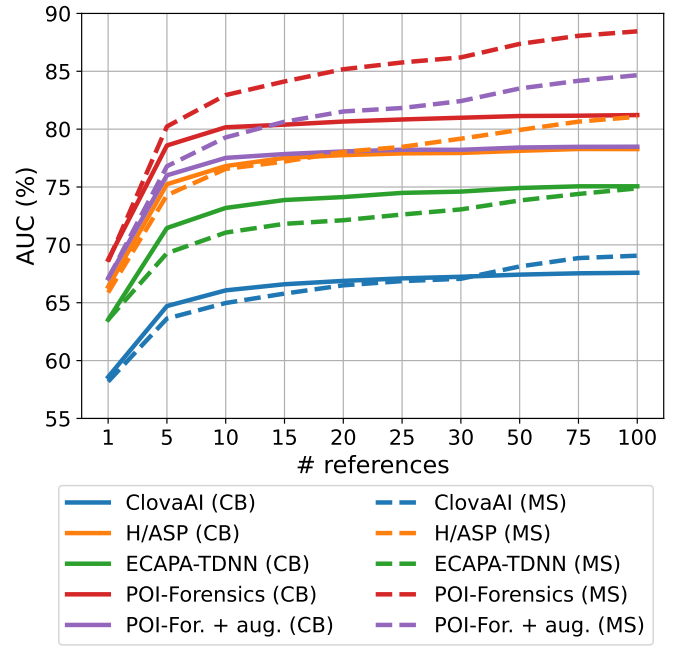


Fig. 2. AUC performance on the IWA dataset as a function of the number of elements in the Reference Set. With the Multi-Similarity testing strategy, the performance keeps growing steadily with no sign of saturation.

footsteps, laughing, mouse-click, keyboard-type, and clock-tick in the indoor scenario and the sounds: engine, train, fireworks, rain, wind, and thunderstorm, in the outdoor scenario. Of course, the performance worsens significantly when noise increases and, in these challenging condition, the Multi-Similarity strategy appears largely preferable. The performance degradation is much reduced when augmentation is included in training. This appears clearly by comparing the red and purple curves, corresponding to POI-Forensics and POI-Forensics with augmentation, using background audios from the MUSAN dataset [23] with random SNR going from 3 to 25dB.

### IV. CONCLUSION

Synthetic audio detection is drawing great attention, lately. However, popular two-class methods, that perform very well on data generated by known methods, generalize poorly to data generated by new unseen methods, proving unfit to deal with real-world scenarios. Here, we propose a new one-class strategy, based on speaker verification tools, which dispenses altogether with fake training audios and therefore provides both a good performance and an excellent generalization ability.

It may seem ironic that speaker verification is used for spoofing detection, when spoofing detection methods were originally conceived to protect speaker verification tools. However, these two approaches fit different scenarios, with different levels of prior information. Our proposal represents just a new tool in the forensic analyst's hands to help counter malicious attacks. However, preliminary results are already very promis-

| | | ASVSpoof2019 | | | FakeAVCeleb | | | IWA | | | average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | EER | t-DCF | AUC | EER | t-DCF | AUC | EER | t-DCF | AUC | EER | t-DCF | AUC |
| Supervised methods | [2] LCNN-LSTM-sum-p2s | 0.03 | 0.08 | 99.4 | 0.39 | 0.84 | 67.3 | 0.33 | 0.99 | 72.1 | 0.25 | 0.64 | 79.6 |
| | [7] RawGATstMul | **0.01** | **0.03** | **99.9** | 0.66 | 1.00 | 30.5 | 0.53 | 0.99 | 47.4 | 0.40 | 0.68 | 59.2 |
| | [9] RawNet2 | 0.21 | 0.70 | 86.1 | 0.55 | 0.99 | 45.5 | 0.51 | 1.00 | 50.7 | 0.42 | 0.89 | 60.8 |
| Centroid-Based testing | [16] ClovaAI | 0.22 | 0.65 | 85.8 | 0.42 | 1.00 | 60.1 | 0.36 | 0.93 | 67.8 | 0.33 | 0.86 | 71.2 |
| | [17] H/ASP | 0.17 | 0.54 | 90.5 | **0.11** | **0.32** | **94.5** | 0.27 | 0.71 | 78.5 | 0.19 | 0.52 | 87.8 |
| | [18] ECAPA-TDNN | 0.20 | 0.62 | 87.6 | 0.17 | 0.46 | 90.4 | 0.30 | 0.79 | 75.4 | 0.22 | 0.62 | 84.4 |
| | [14] POI-Forensics | 0.07 | 0.21 | 97.6 | 0.22 | 0.61 | 86.1 | 0.25 | 0.61 | 81.4 | 0.18 | 0.48 | 88.4 |
| | [14] POI-Forensics + aug. | 0.14 | 0.40 | 93.2 | 0.25 | 0.67 | 83.1 | 0.27 | 0.69 | 78.6 | 0.22 | 0.59 | 85.0 |
| Multi-similarity testing | [16] ClovaAI | 0.24 | 0.74 | 84.2 | 0.40 | 0.98 | 62.8 | 0.31 | 0.93 | 74.1 | 0.32 | 0.89 | 73.7 |
| | [17] H/ASP | 0.20 | 0.65 | 88.3 | 0.14 | 0.41 | 93.5 | 0.22 | 0.63 | 85.2 | 0.19 | 0.56 | 89.0 |
| | [18] ECAPA-TDNN | 0.24 | 0.74 | 84.8 | 0.21 | 0.61 | 86.2 | 0.27 | 0.81 | 79.2 | 0.24 | 0.72 | 83.4 |
| | [14] POI-Forensics | 0.10 | 0.32 | 96.1 | 0.21 | 0.58 | 85.8 | **0.15** | **0.39** | **91.9** | **0.15** | **0.43** | **91.3** |
| | [14] POI-Forensics + aug. | 0.18 | 0.54 | 90.4 | 0.26 | 0.65 | 82.4 | 0.19 | 0.53 | 88.2 | 0.21 | 0.58 | 87.0 |

TABLE II
RESULTS OF ALL CONSIDERED METHODS ON ALL DATASETS IN TERM OF EER, MINIMUM t-DCF AND AUC. THE SUPERVISED METHODS ARE TRAINED ON THE DEVELOPMENT SET OF ASVSPOOF2019 [29] WHILE, THE SPEAKER VERIFICATION METHOD ARE TRAINED ON THE VOXCELEB2 DATASET [31].
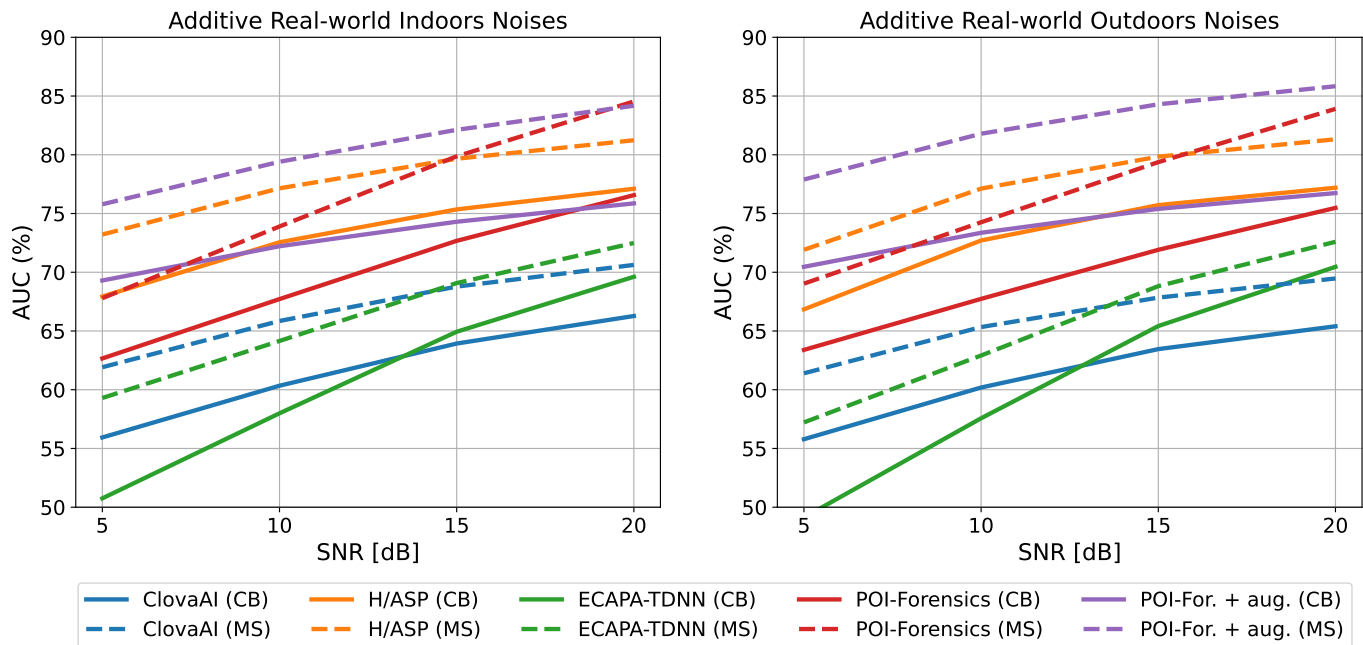


Fig. 3. AUC performance on the IWA dataset in the presence of real noises as a function of the SNR for the indoor (left) and outdoor (right) scenarios. Solid and dashed lines are for the Centroid-Based and Multi-Similarity testing strategies, respectively.

ing, and we hope they will encourage more research on similar solutions.

REFERENCES

[1] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Interspeech*, 2020, pp. 4213–4217.

[2] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[3] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2021, p. 13–22.

[4] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," in *Interspeech*, 2015.

[5] E. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[6] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] 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," *arXiv preprint arXiv:2107.12710*, 2021.

[8] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, pp. 1–14, 2021.

[9] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[10] 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.

[11] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[12] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deep-fake videos from appearance and behavior," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6.

[13] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-Reveal: Identity-aware DeepFake Video Detection," in *IEEE International Conference on Computer Vision (ICCV)*, October 2021.

[14] D. Cozzolino, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," *arXiv preprint arXiv:2204.03083*, 2022.

[15] D. Castan, M. H. Rahman, S. Bakst, C. Cobo-Kroenke, M. McLaren, M. Graciarena, and A. Lawson, "Speaker-targeted synthetic speech detection," in *The Speaker and Language Recognition Workshop*, 2022.

[16] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[17] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[18] 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.

[19] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3652–3656.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[22] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[28] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[29] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[30] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018.

[32] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[33] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[34] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Odyssey*, 2018, pp. 312–319.

[35] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, "Asvspoof 2019: the automatic speaker verification spoofing and countermeasures challenge evaluation plan." 2019. [Online]. Available: http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf

[36] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.