INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speaker-Aware Anti-spoofing

Xuechen Liu[1,2], *Md Sahidullah*[3], *Kong Aik Lee*[4], *Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
[3]Institute for Advancing Intelligence, TCG CREST, India
[4]Institute for Infocomm Research, A*STAR, Singapore

xuechen.liu@uef.fi, sahidullahmd@gmail.com, lee_kong_aik@i2r.a-star.edu.sg, tkinnu@cs.uef.fi

Abstract

We address *speaker-aware anti-spoofing*, where prior knowledge of the target speaker is incorporated into a voice spoofing countermeasure (CM). In contrast to the frequently used speaker-independent solutions, we train the CM in a speaker-conditioned way. As a proof of concept, we consider speaker-aware extension to the state-of-the-art AASIST (audio anti-spoofing using integrated spectro-temporal graph attention networks) model. To this end, we consider two alternative strategies to incorporate target speaker information at the frame and utterance levels, respectively. The experimental results on a custom protocol based on ASVspoof 2019 dataset indicate the efficiency of the speaker information via enrollment: we obtain maximum relative improvements of 25.1% and 11.6% in equal error rate (EER) and minimum tandem detection cost function (t-DCF) over a speaker-independent baseline, respectively.

Index Terms: Speaker Verification, Speaker-Aware Anti-Spoofing, ASVspoof, Deepfake, Spoofing Countermeasures.

1. Introduction

Thanks to recent advances in *neural vocoding* of raw speech waveforms [1], modern *text-to-speech* (TTS) allows the flexible generation of artificial speech that sounds like natural human speech [2]–[4]. Combined with parallel developments in speaker information extraction through *neural speaker embeddings* [5], [6] to condition waveform generation [7], [8], modern TTS allows, in principle, to "put words into anyone's mouth" in the voice of a targeted person.

Despite numerous useful applications, such flexibility raises obvious concerns. First, in the context of biometric authentication, the possibility for an adversary (attacker) to spoof *automatic speaker verification* (ASV) by miscuing oneself as another individual (target) is well known [9]. Second, the potential negative implications of *deepfakes* — a combination of 'deep learning' and 'fake' based on adversarial machine learning [10]— has recently been called to the attention of researchers [11] and the general public [12]. We have already seen alerting examples [13], even if *speech*-related deepfakes have received less attention compared to image- and video-based deepfakes. Deepfakes used for malicious purposes may damage not only the reputation of the targeted individuals but undermine general trust in audio-visual media and biometric technology. To retain the trust, novel protective means are required.

On the positive side, the importance of being able to differentiate "real" inputs from "fake" inputs was proactively recognized early on — way before the concepts of "adversarial machine learning", or "deepfakes" were introduced. In particular, the biometrics research community has studied various *anti-spoofing* methods to protect biometric systems for more than two decades [14]. *Presentation attack detection* (PAD) systems [15], also known as *countermeasures* (CMs), refer to methods aimed at detecting spoofed inputs.

In this study, "CM" refers to a classifier that takes speech input(s) and produces a binary bonafide/spoof prediction. Since 2015, the ASVspoof challenge series [16] has spearheaded benchmarking of speech CMs using common data and performance metrics [17]. Despite its title, the ASVspoof challenges focus on *standalone*, speaker-independent CMs that can be integrated with ASV systems or other applications. Thanks to the common data provided by the ASVspoof challenge and other similar recent initiatives [18], [19], several standalone speech CMs have been developed ranging from early statistical methods [20], [21] to recent deep architectures [22]–[24].

Unfortunately, most existing CMs are far from perfect, particularly when faced with the unknown — be it unseen vocoders, TTS systems, data domains, or codecs [25]. The unconstrained form of the standalone speaker-independent CM task, combined with an artificial speech that is already difficult to differentiate from a real speech by listeners, makes CM generalization beyond training data challenging. The quest for fully general, speaker-independent CM implies that one has to compensate for the potential confounding effects due to speaker, content, and channel variation, with limited prior knowledge.

Even if not addressed in challenges like ASVspoof, in many applications we *do* have prior knowledge of the target person that could readily be utilized by the CM: spoofing attacks are typically targeted against a particular individual — the *same* individual whose identity we seek to verify and who the ASV system already 'knows' based on enrollment data collected earlier. Concerning deepfakes targeted against public figures such as politicians and news anchors, it seems equally safe to assume that we know *who* the intended target in a potential deepfake sample is. For these reasons, it seems then very reasonable to inform CM at the test time of the identity of the hypothesized speaker based on the enrollment sample. To this end, we present an initial investigation on the use of target speaker information for anti-spoofing that we dub *speaker-aware anti-spoofing*.

Our study is not the first one to explore this general idea. The two prior studies [26], [27] that the authors are aware of focus either on replay attack detection with Gaussian mixture model (GMM) backend [26] or on improving the back-end of the ASV system [27]. Our work differs substantially from both studies in terms of CM solutions (statistical model [26] vs. deep learning), the type of fake data (replay attacks [26] vs. synthetic media), the experimental setup, and the evaluation in terms of protocol design and metrics. The main novelty of our work is to propose a precise formulation of the speaker-aware anti-spoofing problem. We also compare different alternative ways

10.21437/Interspeech.2023-1323

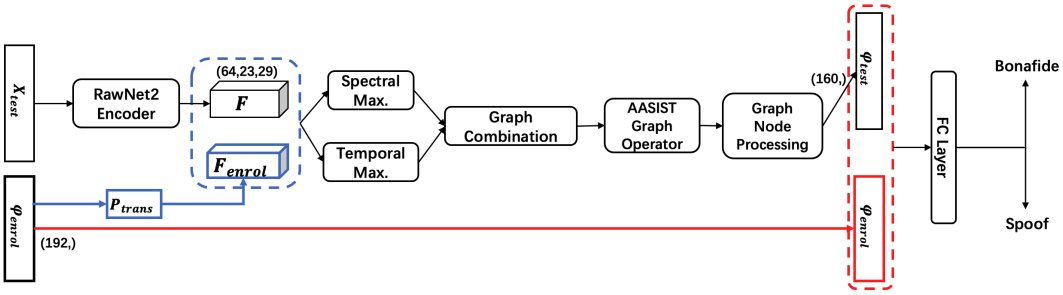

Figure 1: *Illustration of speaker information integration via the enrollment vectors. Blue lines and red lines correspond to different approaches which are not applied simultaneously. Dash lines represent the auxiliary attachment operation. Best viewed in color.*

of integrating target speaker information into the state-of-the-art AASIST model [23]. To be specific, this information is presented using deep speaker embedding and integrated into different parts of AASIST as illustrated in Fig. 1 and detailed in the next section.

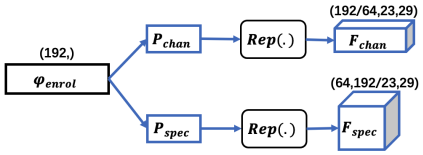

Figure 2: *Illustration of transformation of speaker enrollment vector into channel-wise or spectral-wise attachable components.* $\mathbf{F}_{enr}$ *in Fig. 1 can be either* $\mathbf{F}_{chan}$ *or* $\mathbf{F}_{spec}$*. The transformation matrix* $\mathbf{P}_{trans}$ *can be respectively either* $\mathbf{P}_{chan}$ *or* $\mathbf{P}_{spec}$*.* $Rep(.)$ *denotes repeating operation on the other two dimensions. Best viewed in color.*

2. Speaker-Aware Training of Countermeasure

2.1. Problem Definition

We first define the problem of *speaker-aware anti-spoofing*, as a **binary classification task of discriminating between bonafide and spoofed speech conditioned on the enrolled speaker**. More specifically, it is a conditional hypothesis test defined as follows:

- H_0: Test sample is bonafide and corresponds to the target speaker.
- H_1: Test sample is spoofed and corresponds to the target speaker.

In practice, we address this task by incorporating additional bonafide utterances of the target speaker, detailed next. It is worth noticing that the two hypotheses are conditioned on the target speaker, which make them different from the conventional definition of anti-spoofing.

2.2. Speaker-aware anti-spoofing

The proposed speaker-aware CM is illustrated in Fig. 1. The speaker information can be represented in various ways, from raw audio to well-established deep speaker embeddings. In this study, we focus on the latter. We feed the enrollment audio data into a pre-trained ASV model (here, ECAPA-TDNN [28]). For each enrollment speaker, we extract speaker embeddings correspondingly and average them to get one enrollment embedding: $\varphi_{\text{enrol}} = \frac{1}{N}\sum_{i=1}^{N}\varphi_i$, where N is the number of utterance available for the enrollment.

We consider the recent AASIST [29] for this study. It consists of a speech encoder based on RawNet2 [30]; two heterogeneous graph attention layers operated respectively on spectral and temporal axes; and a graph pooling layer. The pooling layer is followed by node stacking and a fully connected (FC) layer for binary decision-making. As illustrated by the blue lines in Fig. 1, we propose to integrate the enrollment embedding φ_{enrol} into the training by regarding it as auxiliary conditioning information. Methods proposed along with their short-handed forms are presented in the followings.

Integration at the encoder output: Firstly we focus on the output of the encoder, which is a 3-dimensional feature map with channel, spectral, and temporal axes. Let us denote the shape of the map as $(d_{\text{c}}, d_{\text{s}}, d_{\text{t}})$. Inspired by earlier works on channel-wise extension [31]–[33] and speaker adaptation on spectral features [34], we extend our embedding vector $\mathbf{F}_{\text{enr}}$ at either channel-level or spectral-level as illustrated in Fig. 2. The $\mathbf{F}_{\text{enr}}$ in Fig. 1 can thus be either $\mathbf{F}_{\text{chan}}$ or $\mathbf{F}_{\text{spec}}$. $\mathbf{F}_{\text{chan}}$ is of shape $(d_{\text{embed}}, d_{\text{s}}, d_{\text{t}})$ and $\mathbf{F}_{\text{chan}}$ is of shape $(d_{\text{c}}, d_{\text{embed}}, d_{\text{t}})$. Here, d_{embed} is the dimension of the enrollment vector. Such two methods on attaching at channel or spectral axis are denoted as ***enc-chan*** and ***enc-spec***, respectively.

Integration at the encoder output with dimensionality reduction: Since the dimensionalities between the embedding vector and the original feature map are respectively different (d_{embed} vs. d_{s} or d_{t}), one of them may have potentially more impact to model predictions. Therefore, alternatively, we consider including a transformation matrix $\mathbf{P}_{\text{trans}}$ for dimensionality reduction as shown in Fig. 1. $\mathbf{P}_{\text{trans}}$ can be either $\mathbf{P}_{\text{chan}}$ or $\mathbf{P}_{\text{spec}}$, accordingly for channel-level and spectral-level attachment, as illustrated in Fig. 2. In the case where the enrollment vector is firstly reduced to d_{c} or d_{s}, $\mathbf{P}_{\text{trans}}$ is initialized with normal distribution and jointly optimized along with other learnable components in AASIST, and with shape of $(d_{\text{embed}}, d_{\text{c}})$ (for $\mathbf{P}_{\text{chan}}$) or $(d_{\text{embed}}, d_{\text{s}})$ ($\mathbf{P}_{\text{spec}}$) respectively; in the case where dimensionality reduction is not carried out, $\mathbf{P}_{\text{trans}}$ is an identity matrix with shape of $(d_{\text{embed}}, d_{\text{embed}})$. We denote the resulting feature maps by adding the suffix *-reduced*, so the corresponding methods are ***enc-chan-reduced*** and ***enc-spec-reduced***, respectively.

Integration at the FC layer input: Up to this point, we have described the use of enrollment embedding at the early layers of AASIST. As an alternative strategy, we also consider integration before the fully-connected layer, as illustrated in Fig. 1. The input of the FC layer before the decision-making is a utterance-level 160-dimensional vector, denoted as φ_{test}. It has been extracted in earlier works [35] for joint optimization with ASV systems. Here we simply append φ_{enrol} to φ_{test}, with the

input dimension of the FC layer then being $d_{\text{embed}} + 160$. We denote this mean of attachment as ***utterance*** when presenting the results in Section 4.

(a) *Speaker-independent setup*

(b) *Main* setup

(c) *Ablation* setup

Figure 3: *The conceptual illustration of the setups for (a) conventional speaker-independent anti-spoofing, (b) speaker-aware anti-spoofing (Main), (c) ablation study setup (Ablation).* φ_{tar} *represents* φ_{enrol} *from the same speaker of the input audio* $\mathbf{X}_{test}$. φ_{nontar} *represents* φ_{enrol} *from a different speaker from* $\mathbf{X}_{test}$. *Whereas Main complies with the assumption of known target speaker, Ablation is used to assess the impact of the violation of this assumption on CM performance.*

Table 1: *Number of dev/eval trials available for the full CM and customized protocols.*

Setup	dev	eval
Original [36]	24844	71237
Main (Ours)	23780	69252

3. Experimental Setup

Data. We conduct experiments based on ASVspoof 2019 LA [36], which covers 19 types of spoofing attacks. The CM training data consists of 20 speakers (9 male, 11 female) and covers 6 attacks (A01-A06), along with the bonafide condition. The CM evaluation data contains other 13 types of attacks (A07-A19). When training the CM model, for each of the 20 speakers, we generate φ_{enrol} by averaging over his/her speaker embeddings over N randomly selected utterances from the bonafide condition. We set $N = 11$ for female, and $N = 19$ for male.

Protocol. The design logic of the protocol is shown in Fig. 3b, with comparison to the regular anti-spoofing setup shown in Fig. 3a. We recall our assumption that the test utterance originates from a known target speaker, and the task is to determine whether or not the sample is bonafide or spoofed utterance. Therefore, for each test utterance, its associated speaker embedding for enrollment φ_{enrol} is from the same target speaker. In this case, our evaluation protocol is based on the original ASVspoof 2019 CM protocol ("Original" in Table 1), with the bonafide speech trials without the corresponding target speaker in the dataset being removed. We use the speaker information available in the ASV protocol files in the original metadata. The protocol statistics are presented in Table 1. We refer to this protocol setup as *Main*, which differs from the *Ablation* setup described in Section 4.2.

Table 2: *Results in terms of* pooled *EER and minimum tDCF.*

	Main		*Ablation*	
Method	EER(%)	tDCF	EER(%)	tDCF
(Baseline)	1.51	0.043	1.51	0.043
enc-chan	1.48	0.049	1.80	0.061
enc-chan-reduced	2.27	0.077	2.20	0.073
enc-spec	**1.13**	**0.038**	**1.47**	**0.049**
enc-spec-reduced	1.65	0.055	1.88	0.061
utterance	1.89	0.059	1.78	0.052

Model. For the AASIST model, we adopt the solution from the open-sourced repository as the speaker-independent baseline[1]. Model training and hyperparameter setups followed the ones described in [29], except for the batch size that was reduced from 24 down to 12 due to limited computational resources (via a single NVIDIA GeForce GTX 2080Ti). The original shape of the feature map was $(d_{\text{c}}, d_{\text{s}}, d_{\text{t}}) = (64, 23, 29)$, as shown in Fig. 1. We used the open-sourced pre-trained ECAPA-TDNN[2] [28] as the pre-trained ASV model, to extract the speaker embedding with $d_{\text{embed}} = 192$. Each speaker embedding was extracted from the first fully-connected layer after the pooling layer for each input sentence.

Evaluation. We report *equal error rate* (EER) and minimum *tandem detection cost function* (tDCF) [17]. Since compared to minimum tDCF, EER reflects more on the sole CM performance [17], [29], we present our analysis on results primarily on EERs, including the per-attack-type analysis.

4. Results and Analysis

4.1. Results

Results in terms of pooled EER and tDCF are presented in Table 2. While channel-wise speaker integration without dimensionality reduction only marginally improves the EER, the spectral-wise integration works nicely by achieving the lowest numbers in both metrics, outperforming the baseline by relatively 25.1% and 11.6% in terms of EER and minimum tDCF, respectively. This indicates the efficiency of integrating the target speaker integration method on the spectral feature map. The relatively under-performed channel-wise integration, in turn, might be explained by noting the original audio is single channel and contain a rather low level of noise. Applying dimensionality reduction degrades the CM performance for both methods. Attaching the enrollment vector before the FC layer with the bottleneck embedding does not lead to improvements.

A detailed breakdown of the results per attack is shown in Table 3 for the baseline and the best speaker-aware anti-spoofing approach (per protocol). In addition to the CM results shown in the first four lines, the table also displays the EERs of the ASV system on the full CM protocol. These numbers serve to indicate the effectiveness of each attack in spoofing the ASV system (but should not be compared with the CM results). Reflected by the ASV EERs, some attacks do not spoof the ASV model well, such as A09, A17, and A18, which means that those algorithms do not model the speaker information well.

Moving back to the CM performances, there are five types of attacks (A08, A09, A16, A17, A18) where the best-performed proposed method outperforms (or reaches similar performance with) the baseline under both protocols. The ASV EER for four of them is relatively low (lower than 20%) except

[1] https://github.com/clovaai/aasist

[2] https://github.com/TaoRuijie/ECAPA-TDNN/

Table 3: *Results in terms of per-attack-type EER(%) for baselines and best-performed systems. Spoofing attacks in bold font indicate the acquisition of speaker information during the development, according to [36]. The ASV EER is returned by the same pre-trained ECAPA-TDNN model as used in this study, as described in [35].*

	Method	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
Main	(Baseline)	0.75	0.19	0.02	0.88	0.37	0.72	0.14	0.15	0.47	0.73	2.15	4.80	0.78
	enc-spec	1.18	**0.07**	**0.00**	1.38	0.41	0.98	0.22	0.28	0.98	**0.65**	**1.28**	**2.70**	**0.34**
Ablation	*enc-spec*	1.57	**0.08**	**0.00**	1.95	0.47	1.26	0.28	0.35	1.30	**0.71**	**1.79**	**3.13**	**0.30**
ASV EER [35]		32.66	18.80	2.20	50.61	47.08	39.56	11.62	35.39	36.54	60.71	1.85	2.38	4.77

for A16, which indicates that the proposed speaker information integration method further exploits the weakness of the spoofing algorithm by not being able to encode the target speakers well. Improvements can also be observed on A16, which corresponds to the highest ASV EER among all spoofing algorithms. This may exploit the compensation ability of the proposed algorithms on strong attacks that models speaker information well. Future work may further exploit the relationship between the speaker information modeling ability of the spoofing algorithms and its compensation from the CM via such integration.

4.2. Ablation study: Mis-specified speaker identity

The evaluation setup and results described above are based on the assumption that the input audio is target speaker. A natural question that arises is *what might happen if this assumption is violated?* – i.e. how robust the CM is to modeling misspecification in terms of mismatched speaker identities across the enrollment and test utterances. To this end, in this ablation study, we assume that the bonafide input audio is not from its corresponding speaker. In this case, we retain the exact test utterances as in the main protocol but replace the corresponding enrollment utterance with a randomly selected enrollment utterance from another randomly selected speaker. This ablation setup is illustrated in Fig. 3c.

The overall results for the proposed methods for this setup are shown in Table 2. For most proposed methods, compared to the *Main* setup, the results in both metrics are degraded, but not by a large margin. The EER of the best-performed *enc-spec* degrades by relatively 25.1%, but still retains the accuracy of the speaker-independent baseline, even in the severe modeling mis-specification / strong violation of modeling assumption. For *enc-chan-reduced* and *utterance*, the results remain at about same level. The per-attack results for *enc-spec* under this setup is shown in Table 3. For the six types of attacks where improvements are observed under the *Main* setup, *enc-spec* holds its superiority over the baseline, although with marginal performance degradation from *Main* except on A09 and A19. While such degradation indicates the usefulness of target speaker information compared to the one from another speaker, the potential of such *non-target* speaker information still deserves further investigation and extension onto other scenarios.

4.3. Ablation study: Additional bonafide training data

An enrollment vector is not only a speaker representation but also an additional container of bonafide information. Both speaker and bonafide information can be useful as prior conditions for training CM systems. Therefore, we consider an experiment on the effect of additional bonafide training data.

We implement the addition under the full CM protocol by pooling additional speech data from various datasets. We consider VoxCeleb [37] and LibriSpeech [38] corpora. For each dataset, we vary the number of utterances for CM training. The

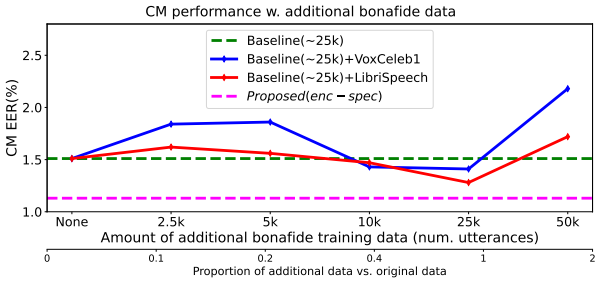

Figure 4: *The relationship between the additional data from different common speech processing datasets and the CM performance under Main. The green dashed line indicates the baseline performance and the pink one indicates the best-performed system. Best viewed in color.*

results are shown in Fig. 4, along with the baseline and the best-performing speaker-aware CM. The figure reveals two interesting patterns. First, the larger amount of additional sole bonafide data from either VoxCeleb1 or LibriSpeech improves performance. Second, the baseline performance is improved with 25k additional utterances, where the amount of data added is almost equal to the total amount of CM training data (25380 utterances [36]). However, adding more data does not necessarily lead to better performance. Relating to the results above, this suggests a more significant benefit provided by additional speaker information, but this might also since the ASVspoof dataset is originated from *VCTK*[3], which is a very clean dataset recorded using the anechoic room. Future work may investigate this issue.

5. Conclusion

We have investigated the feasibility of **speaker-aware anti-spoofing** using state-of-the-art AASIST countermeasure for synthetic spoofing attack detection. Our findings indicate that integration of target speaker enrollment embedding as auxiliary information leads to up to 25.1% relative improvement in anti-spoofing EER. Additional experiments on the effect of alternative speaker information and augmenting the bonafide training using auxiliary corpora have suggested that the proposed speaker-aware training strategy can be more effective. Confirming similar findings done in the two earlier studies using completely different classifiers and datasets [26], [27], this study adds evidence to the positive impact of target speaker prior information. Future work may focus on the Siamese network to encode speaker information and make it available during the training of the CM module, along with more advanced cohort models to encode the speaker information.

[3] https://datashare.is.ed.ac.uk/handle/10283/2651

6. References

[1] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[2] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[3] A. Oord, Y. Li, I. Babuschkin, *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*, PMLR, 2018, pp. 3918–3926.

[4] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, IEEE, 2018, pp. 4779–4783.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[6] D. Snyder, D. Garcia-Romero, G. Sell, *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, IEEE, 2018, pp. 5329–5333.

[7] S. Li, B. Ouyang, L. Li, *et al.*, "Light-tts: Lightweight multi-speaker multi-lingual text-to-speech," in *Proc. ICASSP*, IEEE, 2021, pp. 8383–8387.

[8] M. K. Baskar, L. Burget, S. Watanabe, *et al.*, "Eat: Enhanced asr-tts for self-supervised speech recognition," in *Proc. ICASSP*, IEEE, 2021, pp. 6753–6757.

[9] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, 2020.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, *et al.*, Eds., vol. 27, Curran Associates, Inc., 2014.

[11] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[12] N. Schick, *Deepfakes: The coming infocalypse*. Hachette UK, 2020.

[13] *Top 5 Deepfake Scams That Stormed the Internet This Year*, `https://www.analyticsinsight.net/top-5-deepfake-scams-that-stormed-the-internet-this-year/`, [Online; accessed 10-October-2022], 2022.

[14] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[15] "ISO/IEC 30107. Information technology – biometric presentation attack detection," Standard, 2016.

[16] Z. Wu, J. Yamagishi, T. Kinnunen, *et al.*, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[17] T. Kinnunen, H. Delgado, N. Evans, *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[18] J. Yi, R. Fu, J. Tao, *et al.*, "ADD 2022: The first audio deep synthesis detection challenge," in *Proc. ICASSP*, IEEE, 2022, pp. 9216–9220.

[19] Z. Zhang, Y. Gu, X. Yi, *et al.*, "FMFCC-a: A challenging mandarin dataset for synthetic speech detection," in *International Workshop on Digital Watermarking*, Springer, 2021, pp. 117–131.

[20] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[21] M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015," in *Proc. Interspeech*, 2016, pp. 1700–1704.

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

[23] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP*, IEEE, 2022, pp. 6367–6371.

[24] Z. Teng, Q. Fu, J. White, *et al.*, "ARawNet: A lightweight solution for leveraging raw waveforms in spoof speech detection," in *International Conference on Pattern Recognition (ICPR)*, 2022, pp. 692–698.

[25] T. Chen, A. Kumar, P. Nagarsheth, *et al.*, "Generalization of Audio Deepfake Detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[26] G. Suthokumar, K. Sriskandaraja, V. Sethu, *et al.*, "An analysis of speaker dependent models in replay detection," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[27] D. Castan, M. H. Rahman, S. Bakst, *et al.*, "Speaker-Targeted Synthetic Speech Detection," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 62–69.

[28] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[29] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP*, 2022, pp. 6367–6371.

[30] J.-w. Jung, S.-b. Kim, H.-j. Shim, *et al.*, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms," in *Proc. Interspeech*, 2020, pp. 1496–1500.

[31] D. Cai, X. Qin, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment," in *Proc. Interspeech*, 2019, pp. 4365–4369.

[32] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[33] V. Lostanlen, J. Salamon, M. Cartwright, *et al.*, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[34] V. Gupta, P. Kenny, P. Ouellet, *et al.*, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Proc. ICASSP*, 2014, pp. 6334–6338.

[35] Y. Zhang, G. Zhu, and Z. Duan, "A Probabilistic Fusion Framework for Spoofing Aware Speaker Verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 77–84.

[36] A. Nautsch, X. Wang, N. Evans, *et al.*, "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.

[37] J. Chung, A. Nagrani, E. Coto, *et al.*, "VoxSRC 2019: The first VoxCeleb speaker recognition challenge," in *ISCA archive*, 2019.

[38] V. Panayotov, G. Chen, D. Povey, *et al.*, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.