Continual Audio Deepfake Detection via Universal Adversarial Perturbation

Wangjie Li*, Lin Li* and Qingyang Hong†
* School of Electronic Science and Engineering, Xiamen University, China
† School of Informatics, Xiamen University, China
E-mail: liwangjie@stu.xmu.edu.cn, {lilin, qyhong}@xmu.edu.cn

***Abstract*—The rapid advancement of speech synthesis and voice conversion technologies has raised significant security concerns in multimedia forensics. Although current detection models demonstrate impressive performance, they struggle to maintain effectiveness against constantly evolving deepfake attacks. Additionally, continually fine-tuning these models using historical training data incurs substantial computational and storage costs. To address these limitations, we propose a novel framework that incorporates Universal Adversarial Perturbation (UAP) into audio deepfake detection, enabling models to retain knowledge of historical spoofing distribution without direct access to past data. Our method integrates UAP seamlessly with pre-trained self-supervised audio models during fine-tuning. Extensive experiments validate the effectiveness of our approach, showcasing its potential as an efficient solution for continual learning in audio deepfake detection.**

I. Introduction

In recent years, significant advancements in speech synthesis (TTS) and voice conversion (VC) technologies have made it increasingly difficult to distinguish between genuine and artificially generated speech [1]–[3]. These synthesized voices are often highly realistic, capable of deceiving human listeners with ease. While such technologies offer numerous beneficial applications, they also introduce severe security risks, including privacy violations, identity fraud and other malicious activities. As a result, there is a pressing need to advance audio deepfake detection to keep pace with these evolving threats. Community-driven initiatives, such as the ASVspoof Challenges [4]–[8] and the Audio Deepfake Detection Challenges [9], [10], have played a pivotal role in driving progress in this field. Various techniques, including data augmentation[11]–[13], and multi-feature fusion[14]–[16], have been explored to enhance model generalization by extracting robust audio representations. Furthermore, fine-tuning pre-trained self-supervised learning models has significantly improved detection performance, achieving remarkable results on publicly available datasets [17]–[19].

As speech generation techniques continue to advance, detection models must evolve accordingly to counter emerging spoofing methods. Simply updating the model with newly collected spoofing data risks catastrophic forgetting, where previously learned patterns are partially lost. A straightforward approach to mitigate this issue is to retrain the model using both newly acquired and historical data. Nevertheless, this strategy is not only computationally and storage-intensive but also raises security concerns, such as data leakage. A more effective solution lies in continual learning, which enables the model to adapt to novel attack techniques while retaining knowledge from previous datasets. In the field of audio deepfake detection, few strategies about continual learning have been explored [20]–[22], and most existing approaches employ trainable gradient correction mechanisms to optimize model weights. Although these methods demonstrate effectiveness in mitigating catastrophic forgetting, applying gradient modifications to all neurons imposes constraints on the learning plasticity, potentially limiting its ability to adapt to new spoofing attacks.

In this paper, we propose a novel framework that leverages Universal Adversarial Perturbation (UAP) to preserve historical knowledge in audio deepfake detection. UAP is a distinct type of imperceptible perturbation, specifically crafted to mislead deep learning models with high success rates [23]. Originally introduced for image-related tasks [24]–[26], UAP can be regarded as a feature that captures the primary data-space direction across class boundaries. Given that the distinctions between bona fide and spoofed audio are often subtle, UAP has the potential to serve as a key discriminative feature of spoofed audio relative to real audio. Our method leverages UAP and relevant data to approximate prior feature distributions and integrates them during fine-tuning stage. Instead of storing redundant spoofed data, our method requires retaining only a single UAP generated from the historical model. The combination of UAP and bona fide samples acts as pseudo-spoofed samples, effectively maintaining the distribution of the spoofed class without direct access to prior datasets. Our main contributions are as follows:

- We propose an effective training framework that integrates UAP into fine-tuning the pre-trained self-supervised audio model to preserve historical knowledge. To the best of our knowledge, we are the first to explore the applicability of UAP in audio deepfake detection.
- We investigate the optimal strategy for incorporating UAP in continual learning and demonstrate that feature-level UAP outperforms waveform-level UAP with better retention of prior knowledge.
- Experiments and visualizations validate the effectiveness of our approach, highlighting its potential as a robust and efficient solution for continual audio deepfake detection.

979-8-3315-7206-8/25/$31.00

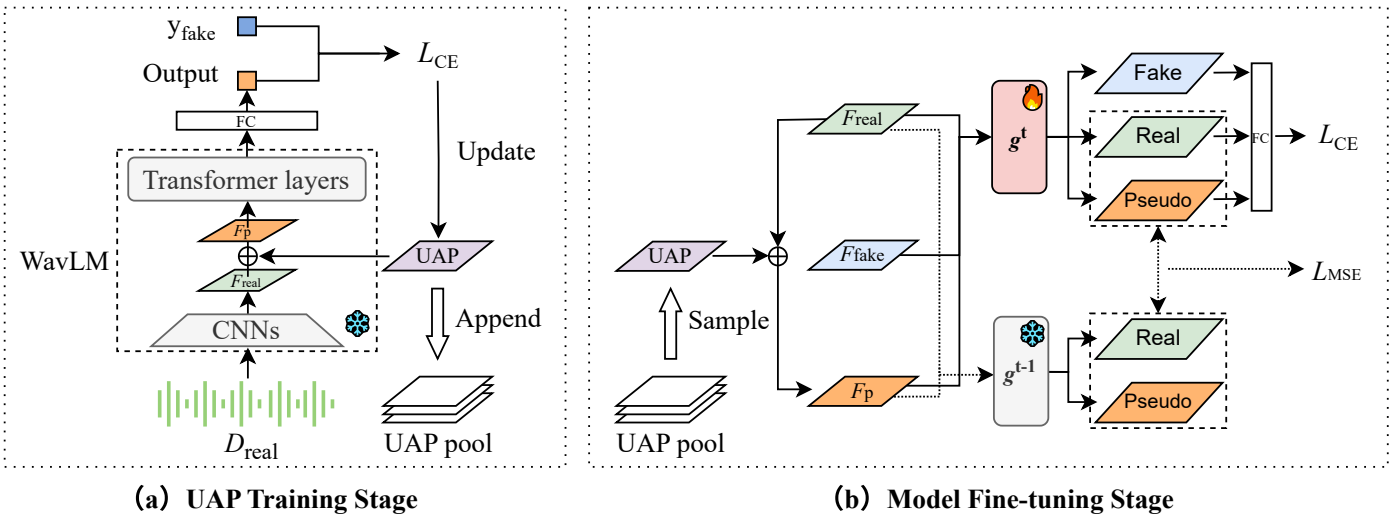

Fig. 1. Overview of proposed training framework for continual audio deepfake detection. Subfigure (a) details the UAP training stage and subfigure (b) elucidates the model fine-tuning stage.

II. Method

A. Base training mechanism

Continual audio deepfake detection aims to create a unified detector capable of handling a sequence of audio data from diverse attack types and domains. Formally, we denote training data available at the t-th stage as $D^t = \{D^t_{\text{real}}, D^t_{\text{fake}}\}$, where D^t_{real} and D^t_{fake} represent the bona fide and spoofed audio respectively. The audio deepfake detection system for the t-th stage is trained exclusively from dataset D^t. Data from the previous stages is no longer accessible. We introduce pre-trained self-supervised audio model to extract general acoustic features $F \subset \mathbb{R}^{\mathbb{B}\times\mathbb{T}\times\mathbb{D}}$ by the frozen convolutional neural network (CNN) layers. The extracted features consist of a batch of frame-level embeddings over time and frequency. The detection model g, consisting of Transformer layers followed by a fully connected layer, is optimized as a binary classifier using cross-entropy loss on dataset D, which is expressed as:

$$L_{CE} = -\frac{1}{|D|}\sum_{x\in F, y\in D} y\log(g(x)) + (1-y)\log(1-g(x)) \quad (1)$$

where $x \in \mathbb{R}^{\mathbb{T}\times\mathbb{D}}$ and $y \in \{0, 1\}$ denote the extracted audio feature and corresponding label.

B. UAP training stage

As illustrated in Figure 1 (a), we generate UAP using the model trained during the $t-1$ phase, and append it to a UAP pool. For the classifier g^{t-1} and bona fide audio feature $x^{t-1}_{\text{real}} \in F^{t-1}_{\text{real}}$ extracted from bona fide subset D^{t-1}_{real} with label $y_{\text{real}} = 0$, our goal is to ascertain a perturbation vector $p^{t-1} \in \mathbb{R}^{\mathbb{T}\times\mathbb{D}}$ capable of misleading the classifier into misclassifying real samples as fake. We target the vector p^{t-1} that conforms to:

$$\begin{aligned} &\text{Pred}(g^{t-1}(x^{t-1}_{\text{real}} + p^{t-1})) = y_{\text{fake}} \\ &\textbf{s.t.} \quad ||p^{t-1}||_\infty \leq \epsilon \end{aligned} \quad (2)$$

where Pred() converts the logit into a prediction value and adjusts the magnitude of the perturbation vector. The pseudo feature x^{t-1}_p is represented as:

$$x^{t-1}_p = x^{t-1}_{\text{real}} + p^{t-1} \quad (3)$$

We utilize the gradient descent technique to refine p^{t-1} by minimizing the entropy between the pseudo feature x^{t-1}_p and the fake label $y_{\text{fake}} = 1$, which is defined as:

$$p^{t-1} = p^{t-1} - \alpha * \text{sgn}(\nabla_{p^{t-1}} \log(g^{t-1}(x^{t-1}_p))) \quad (4)$$

where sgn() represents the symbolic function and α signifies the learning rate of p^{t-1}. The perturbation is reiterated until x^{t-1}_p transgresses the decision boundary. The algorithm concludes when the number of successfully modified prediction outcomes surpasses the predetermined threshold σ. After obtaining p^{t-1}, we integrate it into the UAP pool in preparation for the next model fine-tuning phase.

C. Model fine-tuning stage

Figure 1 (b) demonstrates the integration of the UAP into the model fine-tuning stage. When fine-tuning the pre-tained audio model with newly collected dataset D^t at the t-th phase, UAPs from previous stages up to $t-1$, denoted as $\{p^n\}^{t-1}_{n=1}$, are randomly sampled from the UAP pool at the start of each training iteration. Subsequently, pseudo features F^n_p are created by adding p^n to F^t_{real} on an element-wise basis, formulated as $F^n_p = p^n + F^t_{real}$, and assigned the label $y_{fake} = 1$. Then we combine F^n_{p} with F^t_{real} and F^t_{fake} to fine-tune the classifier.

As parameters are dynamically optimized, the consistency of the pseudo-spoofed distribution may be compromised. To address this, we implement knowledge distillation between prior and current classifiers on the last Transformer layer. This

maintains distillation based on the pseudo-spoofed feature x_p^t, which is formalized as:

$$L_p^t = \frac{1}{|F_p^n|} \sum_{x_p \in F_p^n} ||g^t(x_p) - g^{t-1}(x_p)||^2 \quad (5)$$

In real-world scenarios, bona fide audio tends to be more uniform than spoofed samples [27]–[29]. However, emerging spoofing methods can alter the distribution, impacting the generation of pseudo-spoofed samples. To mitigate this, we apply feature-based knowledge distillation, following the same paradigm as L_p. The previous classifier g^{t-1} serves as the teacher, with a mean-squared loss function used as the regularization term. This can be articulated as:

$$L_r^t = \sum_{x \in F_{real}^t} ||g^t(x) - g^{t-1}(x)||^2 \quad (6)$$

D. Loss function

When new spoofing attacks arise at the t stage, the overall loss function for fine-tuning model with UAP is defined as the sum of the base training loss and aforementioned distribution-preserving optimizations. This can be expressed as:

$$L^t = L_{CE}^t + \lambda(L_{MSE}^t) = L_{CE}^t + \lambda(L_r^t + L_p^t) \quad (7)$$

where λ is a constant determined based on the loss value of L_{CE}^t to ensure they remain at the same order of magnitude.

III. Experimental setup

A. Datasets and metrics

Our experiments are conducted on three publicly available audio deepfake datasets, which are described as follows.

ASVspoof 2019 LA [6] is one of the most commonly used English datasets in audio anti-spoof research. The training and development set share the same attack types including four TTS and two VC algorithms, while the evaluation set contains totally different attack types. The bona fide audio is collected from the VCTK corpus [30].

CFAD [31] is the first public Chinese standard dataset for audio deepfake detection under noisy and transcoding conditions. Twelve mainstream TTS techniques are used to generate spoofed audio. To simulate the real-world scenarios, three noise datasets and six codecs are considered for noise adding and audio transcoding.

ASVspoof 5 [8] is the latest edition of the ASVspoof challenge series. Compared to previous challenges, the ASVspoof 5 database is built upon the MLS English dataset [32], including crowdsourced data collected from a vastly greater number of speakers in diverse acoustic conditions. Attacks are generated and tested using surrogate detection models, while adversarial attacks and contemporary neural codecs are incorporated for the first time.

To investigate the generalization ability of different methods, we also evaluate model performance on ASVspoof 2021 dataset [7]. It is an influential evaluation dataset, sharing the same train and development subsets as ASVspoof 2019.

TABLE I
SUMMARY OF DATA STATISTICS USED IN THIS STUDY. #SPKS, # UTTS, # DUR AND # CONDS REFER TO THE NUMBER OF SPEAKERS, UTTERANCES, APPROXIMATE DURATION AND SPOOFING CONDITIONS, RESPECTIVELY. DIVISION OF TRAIN AND TEST SET IS INDICATED BY / .

Dataset	# Spks	# Utts	# Dur (h)	# Conds
ASVspoof 2019 LA	20 / 48	25,380 / 71,237	48 / 62	6 / 13
CFAD	407 / 728	158,376 / 189,094	176 / 216	8 / 12
ASVspoof 5	400 / 737	182,357 / 680,774	886 / 1345	8 / 16
ASVspoof 2021 LA	- / 48	- / 181,566	- / 130	- / 13
ASVspoof 2021 DF	- / 48	- / 611,829	- / 500	- / 13

Utterances of ASVspoof 2021 LA (21LA) dataset are transmitted over various channels while ASVspoof 2021 DeepFake (21DF) dataset collects numerous utterances processed with various lossy codecs used for media storage. Details are shown in Table I. The Equal Error Rate (EER), which is widely used for audio deepfake detection, is applied to evaluate the experimental performance.

B. Model architecture

We utilize the pre-trained self-supervised audio model WavLM [33] as an efficient feature extractor for audio deepfake detection task. The architecture integrates CNN layers with Transformer encoders to extract speech features at multiple levels. We freeze the CNN layers to extract general acoustic features all the time. After extraction, these features are further processed by twelve Transformer encoders. Each Transformer block features sixteen attention heads and a hidden dimension of 768. Embedding from the last layer is then averaged and passed through the dropout with probability 0.2 and a fully connected layer to get the final prediction score.

C. Implementation details

We applied normalization as pre-processing operation, which is calculated by subtracted mean and divided by standard deviation in each sample independently. Fixed length of four seconds audio segments were used with zero-padding. When evaluating each audio, we make three four-second crops and score them independently. The final score for each sample is obtained by averaging individual scores. To increase the data diversity, we involve noises and reverberation augmented data from the MUSAN and RIR corpus [34], [35], and adopt RawBoost [11] of three noise algorithms. We applied each data augmentation on-the-fly with probability 0.5 during the training process.

We employed the pre-trained WavLM model from HuggingFace[1]. Models were trained using Adam optimizer with hyperparameters $\beta = [0.9, 0.999]$ and weight decay $= 0$. While generating the UAP, we adjust the norm of perturbation (ϵ) to 0.03, set α as 0.0001, and determine the successful threshold (σ) to be 0.8. For fine-tuning process, we set the initial learning rate as $5e-5$ and hyperparameter $\lambda = 5$. We trained the model for 10 epochs with an effective batch size of 128 and selected the best-performing models on the development set.

[1]huggingface.co/microsoft/wavlm-base

TABLE II

THE EVALUATION RESULTS OF UAP WITH DIFFERENT SEQUENCE ORDERS IN TERMS OF EER (%). THE **BEST** RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST ARE UNDERLINED IN EACH EVALUATION SET.

ID	Initial Training Set			Sequence Order	UAP	Evaluation Set			Average
	ASVspoof2019LA	CFAD	ASVspoof 5			ASVspoof2019LA	CFAD	ASVspoof 5	
B1	✓			D_1		2.6	27.3	18.0	16.0
B2		✓		D_2		14.7	5.0	23.1	14.3
B3			✓	D_3		13.9	31.8	3.2	16.3
S1	✓			$D_1 \rightarrow D_2$		8.5	4.8	21.9	11.7
S2	✓			$D_1 \rightarrow D_2$	✓	1.7	**4.3**	19.1	8.3
S3	✓			$D_1 \rightarrow D_2 \rightarrow D_3$		17.3	27.9	**2.8**	16.0
S4	✓			$D_1 \rightarrow D_2 \rightarrow D_3$	✓	2.4	13.0	9.5	8.3
S5			✓	$D_3 \rightarrow D_2$		8.3	4.5	17.0	9.9
S6			✓	$D_3 \rightarrow D_2$	✓	6.2	4.5	13.0	**7.9**
S7			✓	$D_3 \rightarrow D_2 \rightarrow D_1$		**1.5**	16.6	14.6	10.9
S8			✓	$D_3 \rightarrow D_2 \rightarrow D_1$	✓	1.6	13.1	10.7	8.5

Training batches were formed using weighted random sampler with equal probabilities for classes. The complete framework is realized using PyTorch and runs on four NVIDIA A40 GPUs. All results are averaged over three runs.

IV. RESULTS AND ANALYSIS

Since the sequence order is inherently agnostic for continual learning, we evaluate different training orders as shown in Table II. *Order1* follows a progression from simple to complex conditions, denoted as ASVspoof2019LA (D_1) $\rightarrow$ CFAD (D_2) $\rightarrow$ ASVspoof5 (D_3), while *Order2* adopts the reverse order. Both orders involve cross-linguistic adaptation, encompassing a broad range of attack types and pronounced domain gaps.

A. Baseline performance

Table II shows that models trained on a specific dataset achieve high detection performance on their corresponding evaluation sets. However, their ability to generalize to out-of-domain data is significantly limited. Specifically, as seen in baseline B1-B3, models trained on a particular domain (e.g., $D1$) show excellent performance on in-domain evaluation set but fail on evaluation sets from other domains ($D2$ and $D3$). These results validate the hypothesis that models optimized on one dataset struggle to maintain high generalization across diverse and unseen domains.

B. Sequence fine-tuning with UAP

In subsequent experiments, we evaluated the impact of sequence fine-tuning (SFT) following different orders. Our findings indicate that while sequence fine-tuning improves performance on the target domain, it causes catastrophic forgetting, significantly reducing the model's performance on previous domains. In contrast, when UAP is incorporated during model fine-tuning, the model maintains strong performance on previous evaluation sets. Specifically, as shown in S3 and S4, sequence fine-tuning on target domain data results in a marked performance drop on prior domains. While leveraging UAP leads to an average improvement of 48% relative to sequence fine-tuning. This demonstrates that UAP effectively preserves historical learned distribution and mitigates the loss of discriminative knowledge during optimization. Similarly, S7 highlights that sequence fine-tuning, while adapting the model to the target domain, erases knowledge from prior datasets. On the other hand, the utility of UAP in experiment S8 successfully retains the historical spoofing distribution, achieving a 22% relative enhancement in pooled evaluation performance. These findings from both sequence orders validate the effectiveness of our approach for continual audio deepfake detection.

TABLE III

EER (%) COMPARISON OF SYSTEMS WITH DIFFERENT UAP FORMS AFTER SEQUENCE *Order2* ($D_3 \rightarrow D_2 \rightarrow D_1$).

Method	19LA	21LA	21DF	CFAD	ASVspoof5
Joint Training	2.9	3.9	8.6	**4.9**	**4.4**
SFT	**1.5**	4.6	**7.8**	16.6	14.6
UAP(waveform)	1.6	4.3	8.2	14.4	13.8
UAP(feature)	1.6	**3.8**	8.0	13.1	10.7

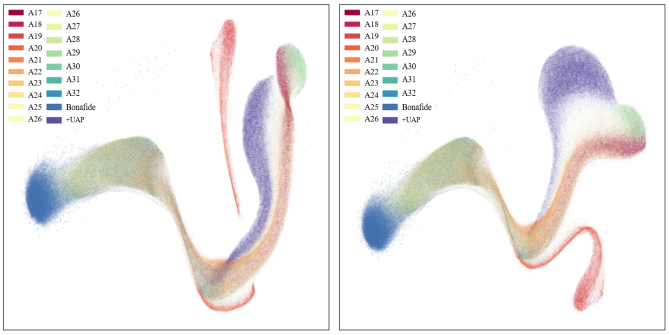

Fig. 2. Visualization of embedding distribution on ASVspoof 5 evaluation set, plotted using UMAP [36] dimension reduction. Left: feature-level UAP. Right: wavform-level UAP.

C. Impact of different UAP forms

In our earlier experiments, we applied perturbations at the feature level to retain the model's previously acquired knowledge. Additionally, we investigated the impact of different forms of UAP during fine-tuning pre-trained models, specifically comparing feature-level and waveform-level UAP.

Our objective was to determine which form of UAP more effectively preserves previously learned distribution while enhancing continual learning performance. Experiments in Table III indicate that although joint training on all datasets achieves strong average performance, it fails to attain optimal performance across all evaluation sets. Nevertheless, leveraging UAP consistently benefits the performance compared to sequence fine-tuning regardless of UAP forms. Moreover, feature-level UAP significantly outperforms waveform-level UAP in retaining prior knowledge of data distributions. This finding highlights the greater potential of feature-level UAP in ensuring better model stability and performance across different domains.

To investigate the underlying reasons for performance discrepancy between two UAP forms, we visualize their effects on the ASVspoof 5 dataset in Figure 2, where $A17$-32 represent eight attack types. The visualization shows that pseudo-spoofed samples created at feature level exhibit a closer distribution with spoofed samples than those at waveform level. We hypothesize that this is because the feature-level UAP captures less redundant details than waveform-level perturbation, making it more sensitive to the model decision. The effectiveness of our method relies on the inherent consistency of bona fide audio. Compared to raw waveforms, features extracted from a pre-trained model exhibit greater uniformity, thereby better preserving historical spoofing distribution and enhancing continual learning stability.

V. Conclusions

In this work, we introduce Universal Adversarial Perturbation (UAP) into audio deepfake detection to retain historical spoofing distribution while adapting to continuously emerging spoofing attacks. Specifically, we propose a training framework that effectively integrates UAP at the feature level when fine-tuning pre-trained self-supervised audio models. Additionally, knowledge distillation is adopted to preserve the distributional consistency of bona fide audio across different training stages. Extensive experiments and visualizations validate the effectiveness of our approach, highlighting its potential as a robust and efficient solution for continual learning in audio deepfake detection. In the future, we plan to explore more diverse adversarial perturbation strategies to better maintain model performance under complex and heterogeneous conditions in audio anti-spoofing.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62371407 and 62276220, and the Innovation of Policing Science and Technology, Fujian province (Grant number: 2024Y0068).

References

[1] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[2] H. Liu, Z. Chen, Y. Yuan, *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning*, PMLR, 2023, pp. 21 450–21 474.

[3] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] 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.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech 2017*, 2017, pp. 2–6.

[6] M. Todisco, X. Wang, V. Vestman, *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, 2019, pp. 1008–1012.

[7] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[8] X. Wang, H. Delgado, H. Tak, *et al.*, "Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *Computer Speech & Language*, p. 101 825, 2025.

[9] J. Yi, R. Fu, J. Tao, *et al.*, "Add 2022: The first audio deep synthesis detection challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9216–9220.

[10] J. Yi, J. Tao, R. Fu, *et al.*, "Add 2023: The second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

[11] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6382–6386.

[12] L. Wang, L. Yu, Y. Zhang, and H. Xie, "Generalizable speech spoofing detection against silence trimming with data augmentation and multi-task meta-learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[13] L. Zhang, K. A. Lee, L. Zhang, L. Wang, and B. Niu, "Cpaug: Refining copy-paste augmentation for speech anti-spoofing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 10 996–11 000.

[14] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," in *Interspeech 2023*, 2023, pp. 3844–3848.

[15] A. Guragain, T. Liu, Z. Pan, H. B. Sailor, and Q. Wang, "Speech foundation model ensembles for the controlled singing voice deepfake detection (ctrsvdd) challenge 2024," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 774–781.

[16] Z. Pan, T. Liu, H. B. Sailor, and Q. Wang, "Attentive merging of hidden embeddings from pre-trained speech model for anti-spoofing detection," in *Interspeech 2024*, 2024, pp. 2090–2094.

[17] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[18] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022.

[19] H. Wu, W. Guo, Z. Zhang, *et al.*, "Spoofing speech detection by modeling local spectro-temporal and long-term dependency," in *Proc. Interspeech 2024*, 2024, pp. 507–511.

[20] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection," in *Interspeech 2021*, 2021, pp. 886–890.

[21] X. Zhang, J. Yi, J. Tao, C. Wang, and C. Y. Zhang, "Do you remember? overcoming catastrophic forgetting for fake audio detection," in *International Conference on Machine Learning*, PMLR, 2023, pp. 41 819–41 831.

[22] X. Zhang, J. Yi, C. Wang, C. Y. Zhang, S. Zeng, and J. Tao, "What to remember: Self-adaptive continual learning for audio deepfake detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 19 569–19 577.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *CVPR*, 2017, pp. 1765–1773.

[24] S. Jetley, N. Lord, and P. Torr, "With friends like these, who needs adversaries?" *NeurIPS*, vol. 31, 2018.

[25] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in *CVPR*, 2020, pp. 14 521–14 530.

[26] K. Sun, S. Chen, T. Yao, X. Sun, S. Ding, and R. Ji, "Continual face forgery detection via historical distribution preserving," *International Journal of Computer Vision*, pp. 1–18, 2024.

[27] 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.

[28] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, "One-class knowledge distillation for spoofing speech detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 251–11 255.

[29] H. M. Kim, K. Jang, and H. Kim, "One-class learning with adaptive centroid shift for audio deepfake detection," in *Interspeech 2024*, 2024, pp. 4853–4857.

[30] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[31] H. Ma, J. Yi, C. Wang, *et al.*, "Cfad: A chinese dataset for fake audio detection," *Speech Communication*, vol. 164, p. 103 122, 2024.

[32] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, 2020, pp. 2757–2761.

[33] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

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

[35] 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 *ICASSP 2017 - 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[36] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 82, Nov. 2024.