Bona fide Cross Testing Reveals Weak Spot in Audio Deepfake Detection Systems

Chin Yuen Kwok[1,2], *Jia Qi Yip*[2], *Zhen Qiu*[1], *Chi Hung Chi*[1], *Kwok Yan Lam*[1]

[1]Digital Trust Centre, Nanyang Technological University, Singapore
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
kwok0062@e.ntu.edu.sg

Abstract

Audio deepfake detection (ADD) models are commonly evaluated using datasets that combine multiple synthesizers, with performance reported as a single Equal Error Rate (EER). However, this approach disproportionately weights synthesizers with more samples, underrepresenting others and reducing the overall reliability of EER. Additionally, most ADD datasets lack diversity in bona fide speech, often featuring a single environment and speech style (e.g., clean read speech), limiting their ability to simulate real-world conditions. To address these challenges, we propose bona fide cross-testing, a novel evaluation framework that incorporates diverse bona fide datasets and aggregates EERs for more balanced assessments. Our approach improves robustness and interpretability compared to traditional evaluation methods. We benchmark over 150 synthesizers across nine bona fide speech types and release a new dataset to facilitate further research at `https://github.com/cyaaronk/audio_deepfake_eval`.

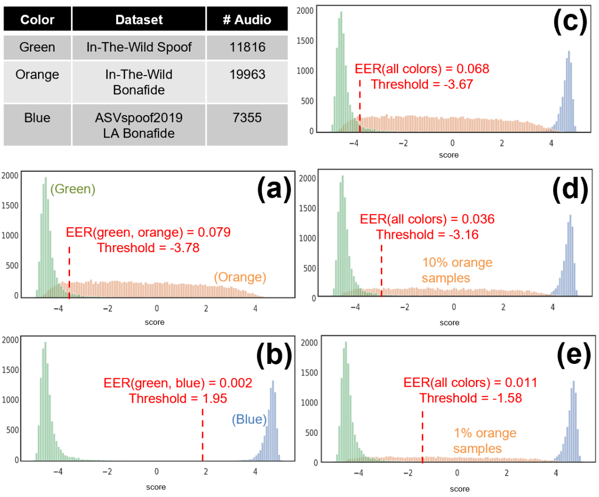
Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Audio deepfake detection (ADD) focuses on identifying synthetic or manipulated audio, commonly referred to as spoofed audio, which aims to replicate genuine recordings. These deepfakes, created using advanced machine learning techniques, pose significant risks to systems relying on voice-based authentication, media forensics, and public trust, especially in the context of misinformation campaigns. The development of robust ADD models is thus critical to mitigate these threats.

Currently, ADD models are primarily evaluated on a single dataset that combines multiple synthesizers [1–3] or through spoof cross-testing [4–6], a method where bona fide audio is paired with M spoof data subsets (one subset corresponds to one synthesizer) to generate M test sets. For each test set, performance is typically reported in terms of Equal Error Rates (EERs). However, this evaluation approach suffers from two key limitations.

First, the framework fails to reflect real-world diversity. Most evaluations are limited to a single bona fide speech type [3,4,6], often focusing on a single environment and speech style (e.g., clean read speech), ignoring the variability present in real-world audio, such as noisy telephony environments, conversational speech, or regional accents. An obvious solution involves merging ADD datasets with diverse bona fide audio datasets. However, this introduces another challenge: as shown in Fig. 1, underrepresented subsets in combined datasets exert less influence on the overall EER, as the EER threshold balances false positives and false negatives across the entire dataset rather than accounting for individual subsets [7].



Color	Dataset	# Audio
Green	In-The-Wild Spoof	11816
Orange	In-The-Wild Bonafide	19963
Blue	ASVspoof2019 LA Bonafide	7355

Figure 1: *Score distributions from three widely used data subsets. The statistics of the subsets are shown in the top-left table. a) The EER for the green and orange subsets uses threshold* -3.78*. b) If the blue subset is used instead of the orange one, the threshold is* 1.95*. c) If the orange and blue subsets are combined, the EER uses threshold* -3.67*. If only d) 10% or e) 1% of the orange subset is included in the test set, the threshold is shifted more towards* 1.95*. This shows that underrepresented subsets in combined datasets may exert less influence on the EER threshold.*

Second, the evaluation framework lacks interpretability. With the proliferation of audio synthesizers that numbers in the thousands[1], the potential number of test datasets is vast. Spoof cross-testing, which reports EERs for every synthesizer dataset, is infeasible and unwieldy, leading to redundancy and reduced clarity. A common workaround is to evaluate on a small subset of recently published or widely used synthesizer datasets [8–12]. However, this approach introduces biases: models that perform well on recent or popular datasets may underperform on older or underrepresented ones [13], leaving potential vulnerabilities undetected. **More critically, the framework does not provide insight into subset-level performance**, specifically whether errors arise from some bona fide types being misclassified as spoof (false positives) or from some spoof types going undetected (false negatives). This lack of granularity makes

[1] `https://huggingface.co/models?pipeline_tag=text-to-speech`

it impossible to accurately diagnose the sources of prediction errors, hindering efforts to improve ADD models.

To overcome these limitations, we propose bona fide cross-testing, a novel evaluation framework that complements spoof cross-testing. Bona fide cross-testing evaluates ADD models using K diverse bona fide speech types, sourced from various datasets, including those not specifically designed for ADD. Each synthesizer testset is paired with the K bona fide types to generate K distinct test sets. This approach ensures that EERs are calculated separately for each bona fide type, addressing the imbalance issue inherent in combined datasets. The process is repeated for each of the M synthesizer testset, yielding $M \times K$ EERs.

Given the large number of available audio synthesizers ($M > 1000$), we propose maximum pooling to summarize EER results, an idea related to Maximum Error Modeling [14]. Specifically, we aggregate the EERs accross all synthesizers and only report the K-highest EERs that corresponds to the most challenging synthesizer testset. This reflects the realistic assumption that attackers will exploit synthesizers that are the hardest to detect [15]. By incorporating both diversity and interpretability, bona fide cross-testing reveals vulnerabilities in ADD models that traditional evaluation methods often overlook.

Our contributions are four-fold:

1. We show that combining multiple data subsets into a single test set is suboptimal because underrepresented subsets may have less influence on the EER threshold. Despite this limitation, previous work [9,12,16] still merges synthesizer subsets into a single test set (e.g. ASVspoof2021-DF [2]), although the subsets vary in size.
2. We reveal that bona fide cross-testing exposes vulnerabilities in ADD models that are often overlooked by conventional evaluation methods such as spoof cross-testing. Notably, bona fide speech types contribute significantly to the challenges faced by ADD systems, whereas traditional approaches primarily focus on the synthesizer type.
3. We propose a novel evaluation framework for ADD that integrates bona fide cross-testing with spoof cross-testing and aggregates EER across spoof tests. Our approach demonstrates greater robustness and interpretability compared to traditional evaluation methods.
4. We provide codes, datasets, and score files required to replicate benchmark results for over 150 synthesizers and 9 bona fide speech types. To the best of our knowledge, this is the first ADD evaluation framework that compares the performance of different bona fide speech types where both the acoustic environment and speech style varies.

2. Methodoglogy

This section begins by presenting the widely used spoof cross-testing evaluation framework. We then propose an extension to this framework, incorporating bona fide cross-testing, and provide justifications for the modifications on the original evaluation approach based on the identified limitations.

2.1. Two types of errors: false positive and false negative

Following the ISO/IEC standard [17] and Zhang et al. [18], we define *spoof* as the positive class and *bona fide* as the negative class. Under this framework, AAD systems are prone to two types of errors: false positives (FP) and false negatives (FN). [2]

- FP (False Positive): the number of *bona fide* audio samples misclassified as *spoof*.
- FN (False Negative): the number of *spoof* audio samples misclassified as *bona fide*.

The normalized (proportional) versions of FP and FN are called the false positive rate (FPR) and false negative rate (FNR), and TP (true positive) and TN (true negative) refer to correctly predicted *spoof* and *bona fide* audio samples, respectively.

2.2. Equal error rate (EER)

EER is a commonly used metric in binary classification tasks. It is a threshold-free metric that represents the error rate at the specific threshold where the FPR is equal to the FNR (or closest to the FNR in the discrete case). FPR $P_{\text{FP}}(\tau)$ and FNR $P_{\text{FN}}(\tau)$ are defined following [18]. In this setting, a nagative class (bona fide) will have a higher classification score and vice versa:

$$P_{\text{FP}}(\tau) = \frac{1}{|\Lambda_{\mathcal{N}}|} \sum_{j \in \Lambda_{\mathcal{N}}} \mathbb{1}(s_j < \tau), \tag{1}$$

$$P_{\text{FN}}(\tau) = \frac{1}{|\Lambda_{\mathcal{P}}|} \sum_{j \in \Lambda_{\mathcal{P}}} \mathbb{1}(s_j \geq \tau), \tag{2}$$

where both $P_{\text{FP}}(\tau)$ and $P_{\text{FN}}(\tau)$ are functions of a predefined threshold τ. $\Lambda_{\mathcal{N}}$ and $\Lambda_{\mathcal{P}}$ are the bona fide and spoof audio subset, respectively. Then, $|\Lambda_{\mathcal{N}}|$ and $|\Lambda_{\mathcal{P}}|$ denote the total number of bona fide and spoof audio samples in each subset. $\mathbb{1}(\cdot)$ denotes the indicator function that outputs 1 when the condition is true and 0 otherwise.

EER is decided by $\hat{\tau}$ where the value of $P_{\text{FP}}(\hat{\tau})$ is infinitesimally close to $P_{\text{FN}}(\hat{\tau})$. Then, EER can be computed by:

$$EER = \frac{P_{\text{FP}}(\hat{\tau}) + P_{\text{FN}}(\hat{\tau})}{2}, \tag{3}$$

where

$$\hat{\tau} = \arg\min_{\tau} |P_{\text{FP}}(\tau) - P_{\text{FN}}(\tau)|. \tag{4}$$

2.3. Spoof cross-testing

Currently, ADD models are primarily evaluated on a single dataset that combines multiple synthesizers [2, 3] or through spoof cross-testing [4–6], a method in which a single bona fide speech type (e.g., clean read speech) is paired with M synthesize datasets (each dataset is only generated by one synthesizer) to generate M test sets. An EER is reported for each test set.

Assume we have one bona fide dataset $\Lambda_{\mathcal{N}}$ and M synthesizer datasets $\{\Lambda_{\mathcal{P}}^m\}_{m=1}^M$, we define the FNR of the m-th spoof subset as:

$$P_{\text{FN}}^m(\tau) = \frac{1}{|\Lambda_{\mathcal{P}}^m|} \sum_{j \in \Lambda_{\mathcal{P}}^m} \mathbb{1}(s_j \geq \tau), \tag{5}$$

Then, the EER of the test set that combines $\Lambda_{\mathcal{N}}$ and the m-th synthesizer dataset $\Lambda_{\mathcal{P}}^m$ is:

$$\text{EER}_m = \frac{P_{\text{FP}}(\hat{\tau}_m) + P_{\text{FN}}^m(\hat{\tau}_m)}{2}, \tag{6}$$

[2] The terms FA and MD used in [19], as well as FR and FA in [20], correspond to FP and FN in this paper, respectively. Although the terminology differs, the definitions remain consistent within the context of the spoofing scenario.

where

$$\hat{\tau}_m = \arg\min_{\tau} |P_{\text{FP}}(\tau) - P^m_{\text{FN}}(\tau)|. \tag{7}$$

2.4. Bona fide cross-testing

As spoof cross-testing is limited to a single bona fide speech type, it ignores the variability present in real-world bona fide audio, such as noisy telephony environments, conversational speech, or regional accents.

To address this, we propose bona fide cross-testing in addition to spoof cross-testing. In addition, as the bona fide speech types in ADD datasets are scarce, we propose to evaluate ADD models using K diverse bona fide speech types collected from various datasets, including those not specific to ADD.

Assume we have K bona fide datasets $\{\Lambda^k_{\mathcal{N}}\}^K_{k=1}$ and M synthesizer datasets $\{\Lambda^m_{\mathcal{P}}\}^M_{m=1}$, we define the FPR of the k-th bona fide dataset as:

$$P^k_{\text{FP}}(\tau) = \frac{1}{|\Lambda^k_{\mathcal{P}}|} \sum_{j \in \Lambda^k_{\mathcal{P}}} \mathbb{1}(s_j \geq \tau), \tag{8}$$

Then, the EER of the test set that combines the k-th bona fide dataset $\Lambda^k_{\mathcal{N}}$ and the m-th synthesizer dataset $\Lambda^m_{\mathcal{P}}$ is:

$$\text{EER}_{k,m} = \frac{P^k_{\text{FP}}(\hat{\tau}_{k,m}) + P^m_{\text{FN}}(\hat{\tau}_{k,m})}{2}, \tag{9}$$

where

$$\hat{\tau}_{k,m} = \arg\min_{\tau} |P^k_{\text{FP}}(\tau) - P^m_{\text{FN}}(\tau)|. \tag{10}$$

2.5. Maximum pooling on spoof cross-testing results

Given the large number of available audio synthesizers ($M > 1000$), we propose average and maximum pooling to summarize the EER results. Specifically, we aggregate the EER results across the M synthesizer types and to report only the K highest and average EERs. The K highest EERs, which correspond to the most challenging synthesizers, are reported to reflect the realistic assumption that attackers will exploit synthesizers that are the hardest to detect.

Our max-pooled EER (mEER$_k$) is defined as:

$$\text{mEER}_k = \max_m \text{EER}_{k,m} \tag{11}$$

We refrain from further aggregating the mEERs across the K bona fide types, as the performance on individual bona fide types is context-dependent and critical. For instance, if an ADD model is deployed to detect fake news, the mEER for bona fide audio in the news domain is more relevant than those for other domains.

3. Experiment Setup

To improve the robustness of the evaluation, we collect audios from more than 150 synthesizers and 9 bona fide audio datasets as shown in Table 1. In-The-Wild [25] is not included in the spoof test subsets as it uses unknown number of synthesizers.

Among the nine bona fide datasets, five of them are sourced from non-ADD corpora, originally designed for ASR [28, 29] and TTS [30] tasks: LibriSpeech test-clean (clean US English read speech) [22], LibriSpeech test-other (noisy and accented speech) [22], AMI IHM (meeting speech with headset microphones) [21], AMI SDM (meeting speech with a single distant microphone) [21], and VCTK (news-domain speech) [23].

As underrepresented subsets in combined datasets may exert less influence on the EER threshold, we further partition the existing ADD datasets such that each data subset is generated by one synthesizer only and EERs are computed separately. To enhance reproducibility, we have released a dataset containing 600 audio samples per synthesizer and bona fide speech type, together with the code to reproduce our bona fide cross-testing evaluation results, available at `https://empty.com`.

Table 1: *Bona fide and spoof test subsets used in our evaluation. The total duration (Dur) in minutes for each subset is reported. SR represents audio sampling rate (in kHz).*

ID	Dataset	Year	SR	Dur	# Audio
	Bona fide test subset (different environments and speech types)				
	AMI-Meeting [21]	2006			
b_1	IHM (meeting)		16	521	13K
b_2	SDM (meeting)		16	521	13k
	LibriSpeech [22]	2015			
b_3	test-clean (storybook)		16	324	2.6K
b_4	test-other (storybook)		16	320	2.9K
b_5	VCTK 0.92 [23] (news)	2019	48	46	755
b_6	FakeAVCeleb-v1.2 [24] (interview)	2021	44.1	1K	10K
b_7	In-The-Wild [25] (social media)	2022	16	1.2K	20K
b_8	EmoFake-EN [6] (emotion)	2022	16	163	3.5K
b_9	AV-Deefake-1M [26] (interview)	2024	44.1	229	1.5K
	Spoof test subset (different synthesizers)				
s_1	ASVspoof2019 LA [5] (13 synthesizers: $s_{1,1}$ - $s_{1,13}$)	2019	16	3.3K	64K
s_2	ASVspoof2021 DF [2] (110 synthesizers: $s_{2,1}$ - $s_{2,110}$)	2021	16	26K	519K
s_3	FakeAVCeleb-v1.2 [26] (1 synthesizer: $s_{3,1}$)	2021	44.1	906	11K
s_4	EmoFake-EN [6] (5 synthesizers: $s_{4,1}$ - $s_{4,5}$)	2022	16	670	14K
s_5	AV-Deefake-1M [26] (4 synthesizers: $s_{5,1}$ - $s_{5,4}$)	2024	44.1	229	1.5K
s_6	CodecFake [4] (6 synthesizers: $s_{6,1}$ - $s_{6,6}$)	2024	16/24	296	4.5K
s_7	MLAAD-v3-EN [3] (19 synthesizers: $s_{7,1}$ - $s_{7,19}$)	2024	22.05	2.5K	19K
s_8	LlamaPartialSpoof [27] (6 synthesizers: $s_{8,1}$ - $s_{8,6}$)	2024	16	6.9K	66K

We evaluate the performance of three recent self-supervised learning (SSL) models, chosen for their robustness across varying audio domains. 1) Wav2Vec-Conformer [9]: Built on the pre-trained XLSR model [31], a variant of wav2vec 2.0, and trained on the ASVspoof2019 LA dataset. 2) Wav2Vec-TCM [12]: An extension of Wav2Vec-Conformer with improved temporal channel dependency modeling. 3) Wav2Vec-SCL [16]: Replaces the Conformer in Wav2Vec-Conformer with linear classifiers and uses Supervised Contrastive Learning (SCL) for learning robust representations. It is trained on ASVspoof2019 LA and additional resynthesized data. The sampling rates are further standardized to match the requirements of each model.

4. Results and Discussions

We begin by presenting the traditional spoof cross-testing evaluation results on 8 ADD test sets. As shown in Figure 2, all models perform well in datasets created in or before 2022.

Table 2: *EER (%) results using our bona fide cross-testing evaluation framework. Given the $M \times K$ EERs obtained from bona fide and spoof cross-testing, we report the K highest and average EERs.*

Method	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	avg.
max. EER of $M = 164$ synthesizers										
Wav2Vec-Conformer [9]	0.95	0.94	0.85	0.86	0.40	0.98	0.93	0.91	0.98	0.87
Wav2Vec-TCM [12]	0.93	0.88	0.83	0.84	0.45	0.98	0.92	0.92	0.99	0.86
Wav2Vec-SCL [16]	0.73	0.90	0.43	0.69	0.31	0.75	0.65	0.57	0.72	0.64
avg. EER of $M = 164$ synthesizers										
Wav2Vec-Conformer [9]	0.11	0.11	0.05	0.08	0.01	0.20	0.12	0.07	0.22	0.11
Wav2Vec-TCM [12]	0.14	0.14	0.04	0.06	0.01	0.15	0.09	0.07	0.15	0.09
Wav2Vec-SCL [16]	0.07	0.14	0.03	0.10	0.01	0.12	0.08	0.04	0.12	0.08

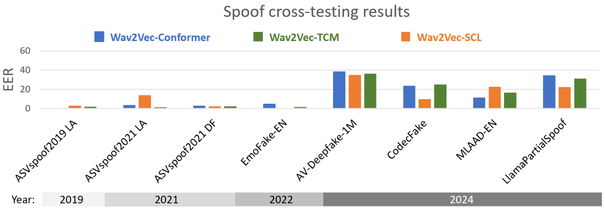


Figure 2: *Traditional evaluation results on 8 ADD testsets [2–6, 26, 27] by combining multiple synthesizers in a single test set. Limitations: (1) Existing ADD datasets typically contain a single bona fide speech type, limiting real-world diversity. (2) Some test sets, such as ASVspoof2021-DF, have an imbalanced distribution of audio samples across synthesizers. Underrepresented subsets may exert less influence on the EER threshold, leading to unfair evaluation. (3) The evaluation does not provide insight into subset-level performance.*

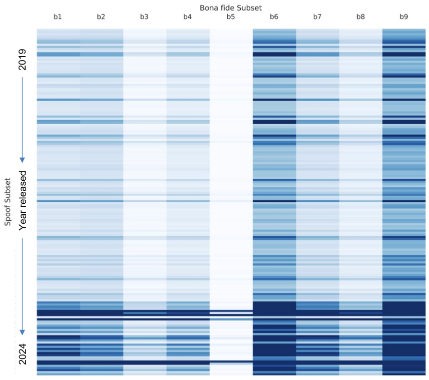


Figure 3: *The results of our bona fide cross-testing combined with spoof cross-testing using Wav2Vec-Conformer [9]. The top-left cell is the EER (darker color means higher EER) of the testset that combines spoof subset $s_{1,1}$ with bona fide subset b_1 as described in Table 1, and the bottom-right cell combines spoof subset $s_{8,6}$ with bona fide subset b_9. The results show that the ADD model performs worse on bona fide subsets b6 and b9 (darker columns), which are celebrity interview speech that may be fast-paced and recorded in a noisy public area.*

However, significant performance degradation is observed for most datasets in 2024, indicating that models trained on older datasets struggle to generalize to newer ones.

The analysis is limited as the evaluation approach has two major limitations: (1) It does not capture real-world diversity, as current ADD datasets include only a limited variety of bona fide speech types. (2) It lacks interpretability, failing to reveal whether errors result from bona fide audio being misclassified as spoof (false positives) or spoof audio going undetected (false negatives).

To address this, we present the results of our bona fide cross-testing combined with spoof cross-testing in Figure 3. The top-left cell is the EER (darker color means higher EER) of the testset that combines spoof subset $s_{1,1}$ with bona fide subset b_1 as described in Table 1, and the bottom-right cell combines spoof subset $s_{8,6}$ with bona fide subset b_9. The results clearly show that the ADD model performs worse on bona fide subsets b_6 and b_9, which are celebrity interview speech that may be fast-paced and recorded in a noisy public area. This shows that the poor performance on certain datasets may stem from the difficulty of detecting bona fide audios recorded in complicated environments.

Then, we aggregate the Equal Error Rates (EERs) across the $M = 164$ synthesizers and report only the $K = 9$ highest and average EERs, as shown in Table 2. The results in the first block indicate that all models fail to detect more than 30% of the spoof audio if only the synthesizer data subset that is the most difficult to synthesizers are considered, The models typically achieve an average EER of approximately 10%. This shows the importance of reporting the maximum EER, as the results show that although current ADD models generally perform well on most synthesizers, they still contain vulnerabilities towards some of the them.

Finally, the results show that among the three models, Wav2Vec-SCL is the most robust against spoof attacks. Specifically, the first block of Table 2 shows that Wav2Vec-SCL generally achieves an EER that is approximately 20% lower than the other models in the worst-case scenario.

5. Conclusion

We introduce bona fide cross-testing, a novel evaluation framework that enhances robustness and interpretability by incorporating diverse bona fide speech types, revealing vulnerabilities overlooked by traditional methods. To support further research, we provide datasets and evaluation codes, aiming to drive the development of more robust and resilient audio deepfake detection models.

6. Acknowledgements

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

7. References

[1] C. Y. Kwok, D.-T. Truong, and J. Q. Yip, "Robust audio deepfake detection using ensemble confidence calibration," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[2] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[3] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," *arXiv preprint arXiv:2401.09512*, 2024.

[4] H. Wu, Y. Tseng, and H.-y. Lee, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," *arXiv preprint arXiv:2406.07237*, 2024.

[5] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. W. D. Evans, M. Sahidullah, V. Vestman, T. H. Kinnunen, K. A. LEE, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, and Z. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, p. 101114, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:211532840

[6] Y. Zhao, J. Yi, J. Tao, C. Wang, and Y. Dong, "Emofake: An initial dataset for emotion fake audio detection," in *China National Conference on Chinese Computational Linguistics*. Springer, 2024, pp. 419–433.

[7] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[8] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 2808–2812.

[9] E. R. Casado, A. G. Alanıs, A. G. Garcıa, A. P. Herreros *et al.*, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Int. Speech Conf.(INTERSPEECH), Dublin, Ireland*, 2023.

[10] H.-j. Shim, J.-w. Jung, and T. Kinnunen, "Multi-dataset co-training with sharpness-aware optimization for audio anti-spoofing," *arXiv preprint arXiv:2305.19953*, 2023.

[11] H.-j. Shim, M. Sahidullah, J.-w. Jung, S. Watanabe, and T. Kinnunen, "Beyond silence: Bias analysis through loss and asymmetric approach in audio anti-spoofing," *arXiv preprint arXiv:2406.17246*, 2024.

[12] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," *arXiv preprint arXiv:2406.17376*, 2024.

[13] B. Chettri, R. G. Hautamäki, M. Sahidullah, and T. Kinnunen, "Data quality as predictor of voice anti-spoofing generalization," *arXiv preprint arXiv:2103.14602*, 2021.

[14] K. Lingasubramanian, S. M. Alam, and S. Bhanja, "Maximum error modeling for fault-tolerant computation using maximum a posteriori (map) hypothesis," *Microelectronics Reliability*, vol. 51, no. 2, pp. 485–501, 2011.

[15] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," *Journal of computer and system sciences*, vol. 80, no. 5, pp. 973–993, 2014.

[16] T.-P. Doan, L. Nguyen-Vu, K. Hong, and S. Jung, "Balance, multiple augmentation, and re-synthesis: A triad training strategy for enhanced audio deepfake detection," in *Proc. Interspeech 2024*, 2024, pp. 2105–2109.

[17] I. J. S. Biometrics, "Iso/iec 30107: Information technology — biometric presentation attack detection," 2016.

[18] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "Range-based equal error rate for spoof localization," 2023. [Online]. Available: https://arxiv.org/abs/2305.17739

[19] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041.

[20] X. Wang and J. Yamagishi, "A practical guide to logical access voice presentation attack detection," in *Frontiers in Fake Media Generation and Detection*. Springer, 2022, pp. 169–214.

[21] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213060286

[24] H. Khalid, S. Tariq, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *ArXiv*, vol. abs/2108.05080, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236976127

[25] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *ArXiv*, vol. abs/2203.16263, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247793039

[26] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," in *ACM Multimedia*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265456085

[27] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and C. E. Siong, "Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation," *ArXiv*, vol. abs/2409.14743, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272827910

[28] C. Y. Kwok, J. Q. Yip, and E. S. Chng, "Continual learning optimizations for auto-regressive decoder of multilingual asr systems," *arXiv preprint arXiv:2407.03645*, 2024.

[29] ——, "Continual learning with embedding layer surgery and task-wise beam search using whisper," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 140–146.

[30] C. Y. Kwok, H. Y. Li, and E. S. Chng, "Asr model adaptation for rare words using synthetic data generated by multiple text-to-speech systems," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1771–1778.

[31] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2006.13979