Unmasking real-world audio deepfakes: A data-centric approach

David Combei[1], *Adriana Stan*[1,2], *Dan Oneata*[2], *Nicolas Müller*[3], *Horia Cucu*[2]

[1]Technical University of Cluj-Napoca, Romania
[2]University "Politehnica" Bucharest, Romania
[3]Fraunhofer AISEC, Germany

david.combei@cs.utcluj.ro, adriana.stan@com.utcluj.ro, dan.oneata@gmail.com, nicolas.mueller@aisec.fraunhofer.de, horia.cucu@upb.ro

Abstract

The growing prevalence of real-world deepfakes presents a critical challenge for existing detection systems, which are often evaluated on datasets collected just for scientific purposes. To address this gap, we introduce a novel dataset of real-world audio deepfakes. Our analysis reveals that these real-world examples pose significant challenges, even for the most performant detection models. Rather than increasing model complexity or exhaustively search for a better alternative, in this work we focus on a data-centric paradigm, employing strategies like dataset curation, pruning, and augmentation to improve model robustness and generalization.

Through these methods, we achieve a 55% relative reduction in EER on the In-the-Wild dataset, reaching an absolute EER of 1.7%, and a 63% reduction on our newly proposed real-world deepfakes dataset, AI4T. These results highlight the transformative potential of data-centric approaches in enhancing deepfake detection for real-world applications. Code and data available at: `https://github.com/davidcombei/AI4T`.

Index Terms: audio deepfake detection, data-centric, real-world deepfakes, antispoofing, data pruning, SSL

1. Introduction

"As you know our FTX exchange is going bankrupt, but you should not panic[...] we have prepared a giveaway. Just go to the site . . . " We are assured by Sam Bankman-Fried (FTX CEO) in a viral clip posted on Twitter in November 2022. Or are we? The video turned out to be fake, but many were fooled by it and, consequently, lost money. Such automatically generated clips (known as deepfakes) are permeating our society resulting in misinformation, scams and making us doubt the veracity of what we are seeing daily on the internet.

Deepfake detection aims to sidestep these risks by automatically identifying synthetically generated contented—video, images, audio. As with any automatic approach, the essential component is data. Data allows training deepfake detection methods, as well as measuring their progress. As such, many datasets are continuously proposed [1–7]. These datasets follow the emergence of recent TTS systems [8–10], or aim to extend the number of speakers or languages [5, 6]. However, these scientific (in the lab) datasets differ in an essential way from real-world deepfakes, i.e. the samples we encounter in the online/social media environment. **Real-world deepfakes** are in most cases released in the online space to mislead. Their development involves a human-in-the-loop approach: the human examines the end result of the speech generator and adjusts its process to obtain an enhanced end-result. The developer would listen to the generated sample and, if needed, update the parameters, input text, or change the generative system altogether–in a wider sense, they would curate one sample at a time. This is in contrast with the **scientific deepfakes** where samples are produced in bulk with minimal additional care towards quality and consistency.

Figure 1: *Fake samples from the proposed AI4T dataset*

To this end, in this paper we contribute a new deepfake dataset, which can be seen as continuation of the work initiated by Müller *et al.* [7]–the first to highlight the importance of in-the-wild data. Our dataset, named AI4T, was collected from YouTube (but some of the samples originated from TikTok, Instagram or Facebook) and it is based on fake videos created for both disinformation and entertainment purposes (see Figure 1). We use only the audio tracks. The dataset totals 13 hours and comprises samples in eight languages.

We find that the performance of state-of-the-art deepfake detection methods on AI4T is well below any other scientific dataset, or even that on the In-the-Wild dataset [7]. We also note that while there is a proliferation of deepfake detection methods, the most performant ones are in essence similar, relying on strong self-supervised learning (SSL) derived features [11–13]. For this reason, instead of over-complicating the model, we take an orthogonal approach and investigate data-centric methods. Data-centric methods have been successfully employed for a wide range of machine learning problems [14–16], yet have received only little attention in the deepfake detection community [17, 18]. We investigate both data- and algorithm-informed sample selection and pruning methods and show that they are a remarkably easy and efficient solution for the challenging out-of-domain, real-world samples.

To summarize, our contributions are as follows: **(i)** we present a novel, diverse and challenging **dataset of real-world audio deepfakes**, curated from online platforms; **(ii)** we demonstrate that state-of-the-art **deepfake detection** systems **struggle with these real-world samples**, exposing a critical performance gap; **(iii)** we show that **data-centric strategies substantially enhance the detection performance** without the need to adapt the underlying model architecture.

Table 1: *An overview of the selected deepfake datasets.*

Dataset short name	Year release	Real data	Langs.	Systems	Real count	Fake count	Duration seconds
ASV19 [19]	2019	VCTK	en	17 TTS/VC	7k	63k	3.1±1.4
FoR [2]	2019	Arctic, LJSpeech, VoxForge, YouTube, TED Talks	en	6 TTS	34k	34k	3.0±2.3
ASV21 [3]	2021	VCTK	en	100 TTS/VC	22k	589k	2.9±1.2
TIM [4]	2023	VidTIMIT	en	12 TTS	430	20k	3.1±1.2
ODSS [5]	2023	VCTK, Hi-Fi TTS, HUI-ACG, SLR-ES	en, es, de	2 TTS	11k	19k	3.1±2.0
MLAAD v5 [6]	2024	M-AILABS	many (38)	82 TTS	80k	154k	8.2±4.5
ASV5 [19]	2024	MultiLingual LibriSpeech	en	13 TTS, 3 VC	50k	183k	9.5±2.3
ITW [7]	2022	YouTube	en	N/A	20k	12k	4.2±3.3
AI4T	2025	YouTube, Instagram, TikTok, Facebook	many (8)	N/A	3k	2k	10.0±0.0

2. Audio deepfake datasets

2.1. Scientific datasets

To address the challenge of detecting real-world deepfakes through data-centric strategies, we start from a wide range of readily available *scientific datasets*. We prioritised datasets in languages that allowed our team of fact-checkers and engineers to conduct subjective evaluations. A comprehensive summary of the dataset pool is provided in Table 1. We note here only the differences with respect to their complete releases. From **Fake or Real (FoR)** [2] we utilize the `for-norm` partition. For **ASV5**, due to its late full release in mid-December 2024, we limited our selection to the training and development subsets alone. From **ASV21**, not wanting to bias the results based on its large size, we randomly selected approximately 60k samples, ensuring the same distribution as the full dataset. To augment the **MLAAD v5** dataset [6] with genuine samples, we selected 80k real samples from the M-AILABS dataset [20].

2.2. Real-world datasets

As baseline real-world samples we start from the **In-the-Wild (ITW)** [7] dataset. It includes a wide variety of content sourced from multiple platforms, environments, and manipulation techniques. The data generation methods of ITW are unknown.

A second real-world dataset is our own contribution and we refer to it as **AI4T**. AI4T dataset includes 196 fake and 192 real videos in 8 languages, with a total duration of around 13 hours. The videos were collected from YouTube, but some of these samples were originally released on other platforms (e.g. Facebook, Instagram or TikTok) over the last two years. The fake samples are diverse in terms of the target audience: political disinformation and financial scams; or harmless entertainment using public and political figures. The samples were identified as fake either based on their metadata or by journalist fact-checkers. The real samples were downloaded from YouTube and maintain similar content to the fakes. To control for length, we segment the original audio into ten-second chunks, resulting in 2005 fake and 2793 real segmented samples. The complete list of audio sources is available in our code repository.

By manually inspecting the samples in these two datasets, we find they lack major artefacts, such as mispronunciations and generation errors. This suggests that the samples have been curated to increase credibility and alignment with a specific goal.

3. Baseline deepfake detection system

We benchmark the detection performance on the newly introduced AI4T dataset. To this end, we build a strong baseline model based on the most promising ideas in deepfake detection. The main ingredient is the use of self-supervised learnt (SSL) audio representations, which have shown strong performance on this task [11, 13, 21]. We then improve this by using intermediary SSL representations and data augmentation.

Table 2: *EER ↓ evaluation for our baseline SSL-based deepfake detection model. ASV19 train+dev subsets are used for training, and rows represent the test data. SotA column reports best results from previous works. (L) refers to the last hidden state, and (B:*) refers to the best SSL layer indicating their number. R is the RawBoost augmentation, C is the codec augmentation and RB+C is the RawBoost plus codec augmentation. Best results are highlighted. Standard deviations are reported only for the columns which involve random selection of data.*

	Dataset	SotA	`xls-r-2b` (L)	 (B:9)	`+RB`	`+C`	`+RB+C`
scientific	ASV19-eval	0.1 [11]	0.6	0.1	**0.07±0.01**	0.1±0.02	0.08±0.0
	FoR	6.9 [21]	6.6	6.6	6.3±0.3	6.4±0.3	**6.0±0.2**
	ASV21	**2.1** [22]	3.3	2.3	2.3±0.1	2.4±0.06	2.4±0.06
	TIM	11.5 [21]	15.2	**5.6**	9.9±1.5	10.9±2.5	12.7±1.5
	ODSS	**16.0** [21]	17.0	16.2	17.2±1.2	16.7±0.4	17.6±0.9
	MLAAD v5	N/A	17.0	12.8	12.6±0.2	**12.4±0.4**	12.9±0.2
	ASV5	N/A	1.7	0.9	0.9±0.06	0.9±0.01	**0.9±0.01**
real	ITW	**3.1** [13]	7.3	3.4	3.5±0.1	3.3±0.2	3.4±0.1
	AI4T	N/A	34.2	**27.4**	27.7±0.4	27.5±0.4	28.0±0.9
	Mean scientific	N/A	8.8	**6.4**	7.0±0.5	7.1±0.5	7.5±0.2
	Mean real	N/A	20.8	**15.4**	15.3±0.3	15.3±0.3	15.7±0.5
	Mean all	N/A	11.5	**8.4**	8.9±0.4	9.0±0.4	9.3±0.2

Initial model. Given an audio file, we extract features from the last hidden state of the frozen wav2vec2 XLS-R 2B model.[1] These features are then average pooled across time yielding a 1,920-dimensional representation. The back-end is a logistic regression classifier trained with weak regularisation ($C = 10^6$). Following prior work [21], we train on the ASV19 training and development sets and report Equal Error Rate (EER). The results in Table 2 show good performance on most of the datasets with the exception of AI4T where the EER is 34.2%.

Intermediary representations. Instead of relying on the last-layer representations, it was showed that intermediary representations provide better features for deepfake detection [13, 23]. We exhaustively explore all layers and find that the best performance on scientific datasets is obtained at layer 9 (out of 48). In Table 2 (column B:9), we observe significant performance improvements over the last layer (column L) for most of the datasets, including ITW and AI4T. However, the performance gap between the two real-world datasets remains.

Data augmentation. To further improve the performance of our baseline, we perform data augmentation [13, 22–28]: we

[1] `https://huggingface.co/facebook/wav2vec2-xls-r-2b`

Table 3: *Dataset mixing EER ↓ results. The table lists the best results for each subset of N datasets along with the baseline.*

# datasets	ASV19	FoR	ASV21	TIMIT	ODSS	MLAAD	ASV5	**ITW**	**AI4T**	**Mean**
baseline	✓							3.41	27.43	15.42
1					✓			4.54	17.24	10.89
2	✓				✓			3.77	14.28	9.03
3		✓			✓	✓		3.07	**13.32**	**8.20**
4		✓		✓		✓	✓	**2.55**	14.14	8.35
5		✓		✓	✓	✓	✓	2.64	13.88	8.26
6	✓		✓	✓	✓	✓	✓	4.96	18.85	11.91
7	✓	✓	✓	✓	✓	✓	✓	2.90	29.03	15.97

augment the training samples using noise (RawBoost [27]) and audio codecs (OPUS and AAC). We partition the dataset into the number of augmentations and apply a single augmentation to each partition. The results in Table 2 (last three columns) show that data augmentation improves the detection for the scientific datasets, but, somewhat surprisingly, no performance gain is observed on the real-world datasets (ITW and AI4T).

These results show that state-of-the-art deepfake detection suffers a significant performance drop when moving from scientific data to the AI4T real-world dataset. Despite this gap, we argue that scientific datasets contain useful information. Next, we search for this information using data-centric approaches, which systematically refine the selection of training samples.

4. Data-centric strategies

Data-centric strategies aim to simplify the computational complexity of a machine learning task through the optimization and enhancement of the data, rather than by modifying the model architecture or algorithm [14, 16, 29]. The idea is that high-quality, well-processed data is often more impactful than complex models in achieving better generalization and accuracy. We do this in the context of audio deepfake detection by exploiting the large number of scientific datasets through two main approaches: at dataset- and at individual sample-level.

4.1. Scientific dataset selection

In Section 3 we showed that when it comes to more recent deepfake samples (i.e., the AI4T dataset), the contents of ASV19 is unable to provide a good generalisation performance for the underlying model. However, as seen in Section 2.1, there are many other datasets that could be used for training. In this section we aim to find the combination that yields the best results. We consider the seven scientific datasets and evaluate all 128 (2^7) combinations. For each dataset we use all of its samples, including real samples and those in the test splits. We use layer 9 features and no data augmentation.

Table 3 presents the best results for each combination of N datasets (single dataset, pairs of datasets, etc.); full results are available in the accompanying code repository. We notice that for ITW, the ASV19-based baseline is already at a good classification performance (3.41% EER), and that the best combination of datasets yields a 1% absolute EER improvement down to 2.55% (line 4). However, for the AI4T dataset almost all dataset mixes yield large improvements, with the best result halving the initial EER to 13.32% (line 3 in Table 3 composed of FoR, ODSS and MLAAD datasets).

To further analyse the results, we plot in Figure 2 the performance of all 128 combinations with respect to the number of datasets used for training We observe that the error does not decrease with the number of datasets and that there is a large variance at each level. These observations suggest that the characteristics of the training datasets are critical for good performance, and some of the datasets might even hurt performance. As an example, using all datasets yields an average EER across the two real-world datasets of 15.97% which is very close to using the popular ASV19 dataset (15.42% EER). If we focus only on the combinations that include ASV19 (green squares and red circles), we see that about about half of the results are better and half are worse than the single-dataset baseline. The best single performing dataset is ODSS: ODSS yields an EER of 10.89% on its own, and this performance can be improved to 8.20% when including also MLAAD and FoR. However, most of the combinations that include ODSS (orange diamonds and red circles) have worse performance than the single dataset.

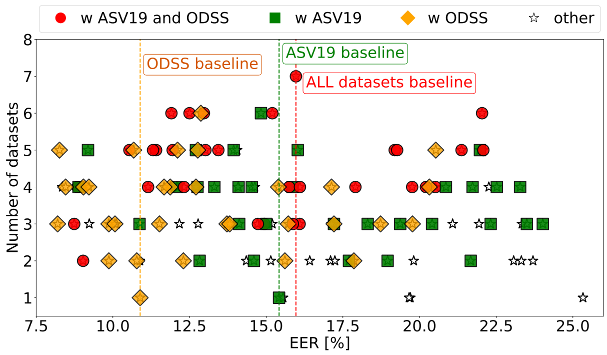


Figure 2: *Average EER ↓ performance for the 2^7 dataset combinations. We use different markers to denote the combinations which include: ASV19 (□), ODSS (◇), or both ASV19 and ODSS (O).*

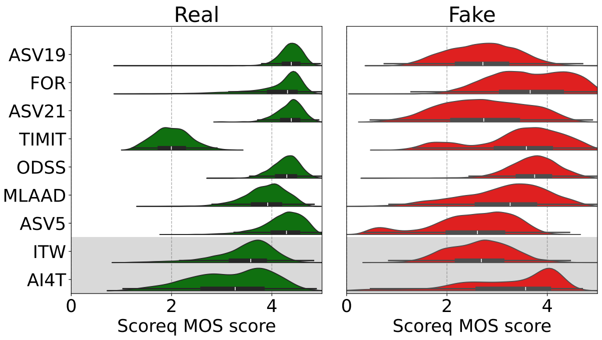
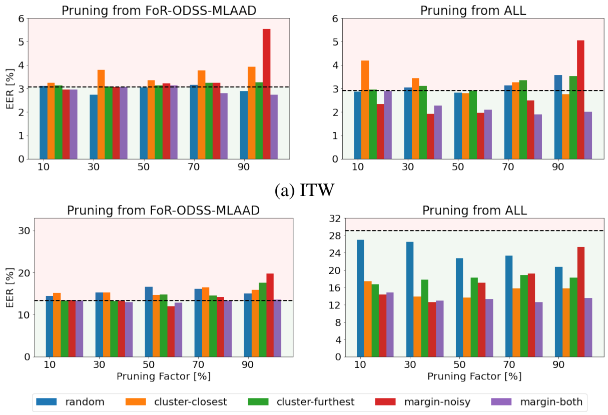


Figure 3: *Distribution plots of the automatic Scoreq MOS scores ↑ over the real and fake samples from all datasets.*

Why do some datasets help, while some other hurt performance? An important aspect is the chronology of these datasets (see Table 1): ITW is closer in time to ASV19, ASV21 and FoR, while AI4T is closer to ODSS, MLAAD and ASV5. The date of release may reflect the dataset quality. We verify this hypothesis by automatically estimating the speech quality with Scoreq [30]. The distributions in Figure 3 indicate that ITW correlates well with the ASV19 and ASV21 data for the fake samples, and with MLAAD (M-AILABS, to be precise) for the real samples. However, AI4T has a much wider score distribution for the real samples than any of the scientific datasets, and a rather high quality for the fake samples–lightly correlated to ODSS and FoR. FoR also contains real data collected from online platforms and fake samples from high-quality commercial speech synthesisers, which may partly explain the increased performance when using it for training.



(a) ITW

(b) AI4T

Figure 4: *Pruning results on the two real-world datasets using five methods, when starting from the best dataset combination (left) or all datasets (right). The horizontal line indicates the performance with no pruning. The pruning factor is the percent of discarded samples. For random pruning we report average results over three random seeds.*

4.2. Sample pruning

Although the deepfake datasets were released as cohesive units, not all their samples may be equally relevant for the generalisation objective. As a result, similar to [17, 31, 32], we explore three data pruning strategies: (i) random pruning; (ii) data-informed pruning; and (iii) algorithm-informed pruning. Random pruning refers to randomly discarding samples from the training data, which was shown to exhibit good generalisation when no sample scoring can used [33, 34]. For data-informed pruning, we rank samples by the distance to their mean, and select either the closest (`cluster-closest`) or the furthest (`cluster-furthest`) samples. The samples are represented by their average pooled self-supervised embeddings and compared using the Euclidean distance. We apply this strategy independently on both real and fake samples from each of N datasets, and obtain $2N$ sets of samples, which we ensemble in the final pruned dataset. For algorithm-informed pruning, we use the logistic regression's margins over the samples. We remove the closest (`margin-noisy`) or the closest and furthest (`margin-both`) points with respect to the decision hyperplane, irrespective of the dataset that they belong to.

Figure 4 shows a subset of the results of the sample pruning methods (complete results are available in the code repository). We use either the best combination of datasets (i.e. `For+ODSS+MLAAD`) or the complete set of available samples (`ALL` data) as the starting pool for the pruning strategies, and report the EER over the real-world datasets. First, we observe that performance is relatively stable with the amount of discarded data. This means that many of the samples carry redundant information. For the data pruned from the FoR+ODSS+MLAAD combination (left side of the figure), the average results obtained after pruning are similar to the baseline. The only pruning method that performs marginally better is `margin-both`. However, obtaining a similar performance when discarding 90% of the data is in itself a very important observation, enabling a more efficient model selection. When pruning the complete set of samples (right side of the figure), margin-based selections work best for both ITW and AI4T. For AI4T, almost all strategies halve the baseline EER (bottom right plot). This may be because the baseline for AI4T is rather poor, at 29.03% EER, and any noisy and outlier samples' reduction are relevant to the logistic regressor.

Table 4: *Data augmentation EER ↓ performance after dataset selection and sample pruning. The augmentation is performed over the samples selected from either the best combination of datasets (For+ODSS+MLAAD), or from the entire set of scientific samples (ALL).*

	FoR+ODSS+MLAAD		**ALL**	
Augm. method	**ITW**	**AI4T**	**ITW**	**AI4T**
No augmentation	2.7	13.5	**1.70**	12.4
+RawBoost	2.5±0.1	11.5±0.3	2.2±0.1	11.7±0.3
+Codecs	2.7±0.3	14.0±0.1	2.0±0.1	12.0±0.7
+RawBoost+Codecs	**2.4±0.4**	**11.3±0.3**	1.9±0.2	**10.2±0.2**

The best results obtained for ITW starting from the two collections of samples are: 2.74% EER for the For-ODSS-MLAAD subset using `margin-both` at 90% pruning factor; and 1.70% EER for the ALL subset using `margin-both` strategy at 80% pruning factor. The corresponding AI4T results are 13.53% and 12.43% EER, respectively. The 1.70% EER obtained for ITW represents a 55% relative increase in performance over the state-of-the-art results, i.e. 3.1% reported by Martin-Donas et al. [13].

4.3. Post-pruning data augmentation

Having selected the best combination of datasets and their most representative samples, we go back to the initial data augmentation strategy in hope of improving the results even further. For all entries reported in Table 4 we adopt the same data augmentation strategy as in Section 3. The *No augmentation* line reports the best results obtained in the previous sections. Given the randomness of the sample augmentation, we report mean and standard deviations over 3 random seeds. As opposed to Table 2 where no gain was observed for the real-world datasets, the results over the pruned datasets show a 2% absolute increase in the performance for the AI4T dataset. There is a relatively limited or no improvement over ITW using this data augmentation strategy, which may be partly caused by the irreducible error within the dataset. These results show that data augmentation can help with generalisation, but only if the underlying data is of sufficient quality.

5. Conclusions

Our results have shown that scientific datasets, while seemingly disjoint from real-world deepfake samples, still contain essential information that can greatly impact model performance. Using a data-centric methodology, we were able to achieve a 55% performance improvement on the ITW dataset (at 1.70% EER) and a 63% performance improvement on the newly proposed and challenging AI4T dataset (at 10.2% EER). While we note that different detection methods may require different selection strategies, we argue that it is essential to prioritize data analysis and data-centric approaches before expanding the capacity of the model (which could potentially obscure its explainability and deployment feasibility).

As future work, we will investigate what makes a relevant sample in the context of deepfake detection. Why are some samples more informative than others? How can we detect the common artefacts of the deepfakes, in order to help us trace them better?

Acknowledgement. This work was co-funded by EU Horizon project AI4TRUST (No. 101070190), and by the Romanian Ministry of Research, Innovation and Digitization project DLT-AI SECSPP (id: PN-IV-P6-6.3-SOL-2024-2-0312), and by the Free State of Bavaria under the DSgenAI project (Grant Nr.: RMF-SG20-3410-2-18-4).

6. References

[1] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 2, 2021.

[2] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *Proc. International Conference on Speech Technology and Human-Computer Dialogue*, 2019.

[3] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, 2023.

[4] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro, "TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection," *IEEE Access*, vol. 11, 2023.

[5] A. Yaroshchuk, C. Papastergiopoulos, L. Cuccovillo, P. Aichroth, K. Votis, and D. Tzovaras, "An Open Dataset of Synthetic Speech," in *Proc. WIFS*, 2023.

[6] 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," in *Proc. IJCNN*, 2024.

[7] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech*, 2022.

[8] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint arXiv:2402.019122*, 2024.

[9] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, and J. Yamagishi, "ZMM-TTS: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, 2023.

[10] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Trans. Assoc. Comput. Linguistics*, vol. 11, 2023.

[11] X. Wang and J. Yamagishi, "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?" in *ICASSP*, 2024.

[12] D. Combei, A. Stan, D. Oneata, and H. Cucu, "WavLM model ensemble for audio deepfake detection," in *Proc. ASVSpoof Workshop*, 2024.

[13] J. M. Martín-Doñas, A. Álvarez, E. Rosello, A. M. Gomez, and A. M. Peinado, "Exploring self-supervised embeddings and synthetic data augmentation for robust audio deepfake detection," in *Proc. Interspeech*, 2024.

[14] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, and X. Hu, "Data-centric AI: Perspectives and challenges," in *Proc. SIAM International Conference on Data Mining*, 2023.

[15] S. Yang, H. Yang, S. Guo, F. Shen, and J. Zhao, "Not all data matters: An end-to-end adaptive dataset pruning framework for enhancing model performance and efficiency," *arXiv preprint arXiv:2312.05599*, 2023.

[16] S. Kumar, S. Datta, V. Singh, S. K. Singh, and R. Sharma, "Opportunities and challenges in data-centric ai," *IEEE Access*, vol. 12, 2024.

[17] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Self-supervised dataset pruning for efficient training in audio anti-spoofing," in *Proc. Interspeech*, 2023.

[18] W. Song, Z. Yan, Y. Lin, T. Yao, C. Chen, S. Chen, Y. Zhao, S. Ding, and B. Li, "A quality-centric framework for generic deepfake detection," *arXiv preprint arXiv:2411.05335*, 2024.

[19] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *Proc. ASVspoof Workshop*, 2024.

[20] The M-AILABS Speech Dataset, "The M-AILABS Speech Dataset," https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/, 2023, accessed on 01/12/2024.

[21] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, "Towards generalisable and calibrated audio deepfake detection with self-supervised representations," in *Proc. Interspeech*, 2024.

[22] 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," in *Proc. Interspeech*, 2024.

[23] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised XLS-R and SLS classifier," in *Proc. ACM Multimedia*, 2024.

[24] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC antispoofing systems for the ASVspoof 2021 challenge," in *Proc. ASVspoof Workshop*, 2021.

[25] T. Chen, E. Khoury, K. Phatak, and G. Sivaraman, "Pindrop labs' submission to the asvspoof 2021 challenge," in *Proc. ASVspoof Workshop*, 2021.

[26] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *Proc. ICASSP*, 2023.

[27] H. Tak, M. R. Kamble, J. Patino, M. Todisco, and N. W. D. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proc. ICASSP*, 2021.

[28] 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 *Proc. Speaker and Language Recognition Workshop*, 2022.

[29] G. Press, "Andrew Ng launches a campaign for data-centric AI," *Forbes*, 2021, accessed: 2025-01-20. [Online]. Available: https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/

[30] A. Ragano, J. Skoglund, and A. Hines, "SCOREQ: Speech quality assessment with contrastive regression," in *Proc. NeurIPS*, 2024.

[31] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Representative subset selection for efficient fine-tuning in self-supervised speech recognition," *arXiv preprint arXiv:2203.09829*, 2022.

[32] M. Lindsey, N. R. Robinson, F. Kubala, and R. M. Stern, "Reducing the cost of spoof detection labeling using mixed-strategy active learning and pretrained models," in *Proc. ASRU*, 2023, pp. 1–7.

[33] F. Ayed and S. Hayou, "Data pruning and neural scaling laws: fundamental limitations of score-based algorithms," *Trans. Mach. Learn. Res.*, 2023.

[34] A. Vysogorets, K. Ahuja, and J. Kempe, "DRop: Distributionally robust pruning," in *Proc. ICLR*, 2024.