EnvSDD: Benchmarking Environmental Sound Deepfake Detection

Han Yin[1,2], Yang Xiao[3,4], Rohan Kumar Das[4], Jisheng Bai[1], Haohe Liu[5]
Wenwu Wang[5], Mark D Plumbley[5]

[1]Northwestern Polytechnical University, China; [2]Speech Lab, Alibaba Group, China; [3]The University of Melbourne, Australia; [4]Fortemedia Singapore, Singapore; [5]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK

yinhan@mail.nwpu.edu.cn

Abstract

Audio generation systems now create very realistic soundscapes that can enhance media production, but also pose potential risks. Several studies have examined deepfakes in speech or singing voice. However, environmental sounds have different characteristics, which may make methods for detecting speech and singing deepfakes less effective for real-world sounds. In addition, existing datasets for environmental sound deepfake detection are limited in scale and audio types. To address this gap, we introduce EnvSDD, the first large-scale curated dataset designed for this task, consisting of 45.25 hours of real and 316.74 hours of fake audio. The test set includes diverse conditions to evaluate the generalizability, such as unseen generation models and unseen datasets. We also propose an audio deepfake detection system, based on a pre-trained audio foundation model. Results on EnvSDD show that our proposed system outperforms the state-of-the-art systems from speech and singing domains.

Index Terms: sound deepfake detection, environmental soundscape, deep learning, EnvSDD

1. Introduction

Imagine sitting in a college classroom when you suddenly hear a fire alarm start, its sound gradually growing louder. At first, you may think it is a false alarm. Then soon, you hear frantic rush of footsteps and even a siren at a distance. Instinctively, most of you would rise from your seats and begin to move toward the exit. But what if these sound events are not real at all? What if these sounds are created by artificial intelligence (AI)? A similar event was shown in a TikTok video[1] that received over 10 million views. This story shows a new and growing concern in our society: the rise of AI-generated environmental sounds. With the growth of open-sourced audio generation models, such as AudioLDM [1] and AudioLCM [2], it is now becoming easier for people to create very realistic environmental sounds, which previously needed entire teams of professional individuals to produce. While these technologies have vast applications in fields such as virtual reality and media production, they also pose serious threats, such as the potential for malicious use in misinformation and the generation of fake audio content that can mislead the public [3]. Therefore, these societal concerns show a strong need to develop methods that can detect fake environmental sounds accurately.

The goal of environmental sound deepfake detection (ESDD) is to determine whether a sound clip is sourced from real-life scenarios or has been artificially generated by models. Recently, speech and singing voice deepfake detection have received widespread attention, leading to various challenges and benchmarks [4–6]. However, ESDD poses challenges that are different from those in speech and singing domains. For example, speech and singing have specific pitch ranges, but environmental sounds lack steady rhythms and fixed tones. In addition, for environmental sounds, many sound events may occur simultaneously, a phenomenon that may not be as apparent in pure speech or singing clips. Therefore, we question whether methods from speech and singing deepfake detection can be directly applied to environmental sounds.

Several studies have focused on fake environmental sound detection. In SceneFake [7], the authors first enhance the real speech involving a scene (e.g., "Airport"), then add another scene to the enhanced speech (e.g., "Street"), resulting in a scene-fake audio clip. Similar to SceneFake, in EnvFake [8], the authors create a multimodal scene-fake dataset by combining visual and audio information. These works focus on scene-consistency detection, where all environmental audio clips are sourced from real-life scenarios rather than being generated by AI models. Ouajdi et al. [9] use real-fake paired audio data from a Foley sound synthesis challenge for ESDD, where fake sounds are generated by 44 submitted audio generation systems. However, this work only includes monophonic audio, and only conducts in-domain evaluation, without considering the models' generalization to unseen domains. In FakeSound [10], the authors mask a portion of the audio and use audio inpainting models to generate new audio in the masked segments, creating fake sound clips. In this work, most audio clips are polyphonic, with limited consideration of monophonic conditions. In addition, they use a relatively small subset of AudioCaps [11] for generating fake audio, with just about 5 hours of data.

In our work, we present EnvSDD, the first large-scale dataset for ESDD, with 45.25 hours of real audio and 316.74 hours of deepfake environmental sound clips. We sample real-life sound clips from six datasets, covering both monophonic and polyphonic conditions. Additionally, we use five state-of-the-art (SOTA) audio generation models to generate the deepfake audio clips. Specifically, we consider two types of deepfakes, including the two primary paradigms of audio generation: text-to-audio (TTA) and audio-to-audio (ATA). To evaluate the dataset, we first train two SOTA systems from the speech and singing voice deepfake detection domains on EnvSDD. Then, we propose a new system leveraging a pre-trained audio foundation model (i.e., BEATs [12]), to improve the detection performance. The main contributions of this paper are as follows.

- We present EnvSDD, a large-scale dataset for ESDD.
- We propose a new detection system, which outperforms the baselines across all evaluation scenarios. The dataset, models, and codes are all publicly available.[2]
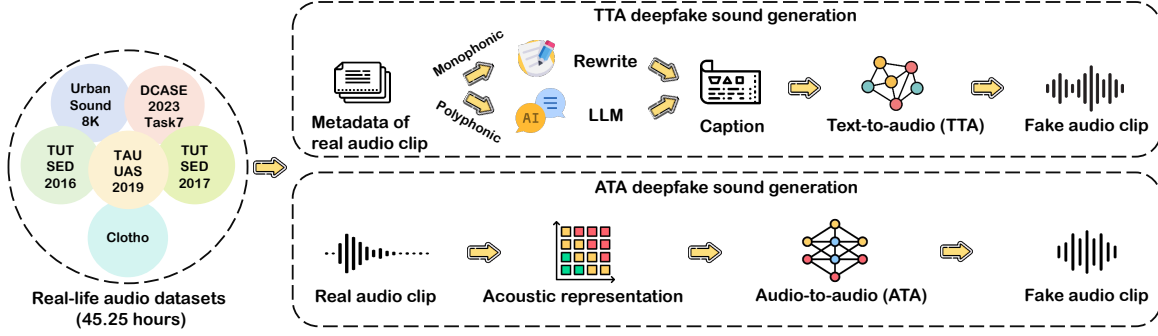

Figure 1: *Pipeline for creation of the proposed EnvSDD dataset.*

2. EnvSDD Dataset

As shown in Figure 1, the EnvSDD dataset includes two types of deepfakes: TTA and ATA. We first sample real-life audio clips from various publicly available datasets and then generate deepfake clips using the TTA and ATA models.

2.1. Real Data Collection

EnvSDD includes both monophonic and polyphonic audio. Monophonic audio contains a single event per clip, while polyphonic refers to audio with multiple overlapping sound events. These two types are considered to account for the diverse range of real-world scenarios, where both isolated and overlapping sound events occur frequently. For monophonic audio, we use UrbanSound8K [13] and DCASE 2023 Task7 Dev [14] as the sources of real data. For polyphonic audio, we utilize TAU UAS 2019 Open Dev [15], TUT SED 2016 [16], TUT SED 2017 [17, 18] and Clotho [19]. All audio samples are segmented into 4-second clips, and since most audio generation models are trained with 16 kHz audio, we resample all clips to 16 kHz.

Monophonic datasets: **UrbanSound8K** is a widely used dataset for environmental sound classification, specifically designed for urban sound recognition tasks, which contains 10 different urban sound classes, such as "siren" and "drilling". We select audio clips with a duration of 4 seconds and a sampling rate above 16 kHz from it, resulting in 4,523 clips. **DCASE 2023 Task7 Dev** refers to the development set of Task 7 from DCASE 2023 challenge, which contains real-life audio clips from various public resources with 7 event classes.

Polyphonic datasets: **TAU UAS 2019 Open Dev** is designed for acoustic scene classification, consisting of audio clips recorded in 14 different real-life acoustic scenes, such as "airport" and "office". **TUT SED 2016** and **TUT SED 2017** are two environmental audio datasets for sound event detection, containing real-life recordings from various indoor and outdoor scenarios. **Clotho** is a dataset designed for audio captioning tasks, which contains audio clips from FreesSound [20], and human-written captions describing the audio content.

Since DCASE 2023 Task 7 Dev and Clotho are used only for evaluation, we select a small subset from these datasets, covering various real-life scenarios. Additionally, the "DogBark" and "GunShot" events in DCASE 2023 Task 7 Dev overlap with UrbanSound8K, and are therefore excluded.

2.2. Deepfake Data Generation

In the proposed EnvSDD dataset, pre-trained TTA and ATA models are applied to generate deepfake audio clips. As shown in Table 1, we use five models for TTA and two models for ATA, with an inference step of 100.

Table 1: *Overview of audio generation models used in the proposed EnvSDD dataset.*

Deepfake Type	Generation Model (G#)
Text-to-Audio	AudioLDM (G1)
	AudioLDM 2 (G2)
	AudioGen (G3)
	TangoFlux (G4)
	AudioLCM (G5)
Audio-to-Audio	AudioLDM (G1)
	AudioLDM 2 (G2)

TTA Deepfake: We used 5 models for generating TTA deepfake clips, namely, AudioGen [21], AudioLDM [1], AudioLDM 2 [22], TangoFlux [23] and AudioLCM [2]. An important challenge is how to generate accurate and meaningful captions for audio clips, which are applied as inputs to the TTA models.

As shown in Figure 1, for monophonic audio, we directly rewrite the metadata as the audio caption. "Rewrite" refers to removing meaningless characters from the metadata, capitalizing the first character, and adding a period in the end (e.g., "gun_shot" is rewritten to "Gun shot."). For polyphonic audio, a large language model (LLM) is applied to generate captions based on the metadata. Specifically, we use Mistral 7B [24] as the LLM, with prompts configured as follows. If the metadata contains both scene and event labels, the prompt is "Prompt A" here as shown in the box:

> ### Prompt A
>
> This clip is recorded in {scene label}, where the following events or sounds are happening: {event labels}.
>
> A caption is a descriptive sentence, which vividly depicts the acoustic content of the audio clip. Please provide one sentence for the caption to directly describe the sound.

If the metadata contains only the scene label, the prompt is:

> ### Prompt B
>
> This clip is an audio clip recorded in {scene label}.
>
> A caption is a descriptive sentence, which vividly depicts the acoustic content of the audio clip. Please provide one sentence for the caption to directly describe the sound that might occur in the scene.

By prompting Mistral 7B, we generate a caption for each audio clip. It is noted that Clotho already provides audio captions, so we do not generate captions for it.

Table 2: *Statistics of the proposed EnvSDD dataset.*

Audio Type	Source Dataset (D#)	#Real	#Fake (TTA)	#Fake (ATA)
Monophonic	UrbanSound8K (D1)	4,523	22,615	9,046
	DCASE 2023 Task7 Dev (D2)	500	2,500	1,000
Polyphonic	TAU UAS 2019 Open Dev (D3)	31,700	158,500	63,400
	TUT SED 2016 (D4)	1,695	8,475	3,390
	TUT SED 2017 (D5)	1,806	9,030	3,612
	Clotho (D6)	500	2,500	1,000

Table 3: *Statistics of train, validation, and test sets in EnvSDD.*

Split	Source Datasets	Generation Model	#Real	#Fake
TTA Deepfake				
Train	D1, D3, D4, D5	G1, G2, G3	27,811	83,433
Valid	D1, D3, D4, D5	G1, G2, G3	7,942	23,826
Test 01	D1, D3, D4, D5	G1, G2, G3	3,971	11,913
Test 02	D1, D3, D4, D5	G4, G5	3,971	7,942
Test 03	D2, D6	G1, G2, G3	1,000	3,000
Test 04	D2, D6	G4, G5	1,000	2,000
ATA Deepfake				
Train	D1, D3, D4, D5	G1	27,811	27,811
Valid	D1, D3, D4, D5	G1	7,942	7,942
Test 01	D1, D3, D4, D5	G1	3,971	3,971
Test 02	D1, D3, D4, D5	G2	3,971	3,971
Test 03	D2, D6	G1	1,000	1,000
Test 04	D2, D6	G2	1,000	1,000

ATA Deepfake: AudioLDM and AudioLDM 2 are used for ATA, where a latent diffusion model conditions on acoustic representations to generate audio events resembling those in the input. AudioLDM relies on CLAP [25] features, while AudioLDM 2 leverages AudioMAE [26], which captures finer acoustic details. Therefore, ATA with AudioLDM 2 produces audio that more closely matches the original, while AudioLDM generates semantically similar but acoustically distinct outputs.

2.3. Dataset Splits

For clarity, as shown in Table 1 and Table 2, we label the source datasets as D1-D6 and the generation models as G1-G5. Table 3 presents statistics for different dataset splits. To comprehensively evaluate detection performance across various scenarios, we define four distinct test conditions.

Test 01 is for in-domain evaluation, where the source datasets and audio generation models are seen during training. Test 02-04 are designed for out-of-domain evaluation. Specifically, Test 02 evaluates generalizability by using generation models that are different from those during training, assessing the detection model's ability to handle unseen audio generation models. Test 03 focuses on generalizability to unseen datasets by ensuring the source datasets are significantly different from those used in training and validation. Test 04 presents the most challenging scenario, as both the source datasets and generation models differ from those during training.

3. Model Architectures

3.1. Baselines: AASIST and W2V2+AASIST

Recent studies in speech deepfake detection show that key clues can appear in both the spectral and temporal domains [27]. AA-

SIST [28] is an end-to-end system that uses a novel heterogeneous stacking graph attention layer [29] to learn these features. This model has been applied as the baseline in various speech and singing voice deepfake detection challenges [4–6]. Building on AASIST, W2V2+AASIST [30] uses a pre-trained speech foundation model, wav2vec 2.0 XLS-R [31], as the front-end, to extract representations from the waveform, achieving SOTA results on ASVspoof2021 [5] and SingFake datasets [6]. The wav2vec 2.0 XLS-R (W2V2) model was pre-trained on 128 languages and approximately 436K hours of unlabeled speech data with self-supervised learning. We apply these two systems as the baselines, to explore whether methods from speech and singing domains can be applied to environmental sounds.

3.2. Proposed BEATs+AASIST System for ESDD

While W2V2+AASIST works well for speech deepfakes and shows the benefits of a pre-trained speech foundation model, we hypothesize that such a model may not generalize well for diverse sound scape in environmental sounds. Therefore, we propose a novel system that integrates an audio foundation model with AASIST. We select the audio foundation model due to its pre-training on large-scale, diverse sound datasets, enabling it to capture robust and rich acoustic features for environmental audio. In our work, we use BEATs [12], a pre-trained model that learns deep audio representations from AudioSet-2M [32] with self-supervised learning. BEATs begins with a random projection as an acoustic tokenizer for mask and label prediction, then refines the tokenizer with its learned knowledge in repeated cycles. As in W2V2+AASIST, the output of BEATs is fed to a RawNet2-based residual encoder [33] that learns higher-level feature representations. Then a self-attention aggregation layer is applied to extract the acoustic representation, which is fed to the AASIST model to obtain a two-class prediction.

By replacing the W2V2 front-end with BEATs, our proposed system is expected to capture the complex nature of environmental sounds more precisely. This integration of BEATs with AASIST is anticipated to enhance performance in detecting environmental audio deepfakes. We evaluate all the three systems on our proposed EnvSDD dataset.

4. Experiments and Results

4.1. Experimental Setup

During training and fine-tuning, we use a batch size of 32 and the Adam optimizer with a weight decay of 0.0001. When training AASIST from scratch, the initial learning rate is 0.001. For fine-tuning W2V2+AASIST and BEATs+AASIST, we reduce the learning rate to 0.00001 to avoid overfitting. The maximum number of epochs is 50. Training stops early if the validation loss does not decrease for 5 consecutive epochs.

We consider equal error rate (EER) [34] as the evaluation metric, which is also common in speech and singing deepfake

Table 4: *Performance in EER (%) of baseline and proposed systems on EnvSDD dataset. "Seen SD" means "Seen Source Datasets" and "Seen GM" means "Seen Generation Model".*

System	Test Set	Test Condition		Fake Type	
		Seen SD	Seen GM	TTA	ATA
AASIST	Test 01	✓	✓	0.66	0.28
	Test 02	✓	✗	3.70	0.68
	Test 03	✗	✓	6.80	3.00
	Test 04	✗	✗	17.50	4.40
	Average	-	-	7.17	2.09
W2V2+AASIST	Test 01	✓	✓	0.26	0.38
	Test 02	✓	✗	13.04	26.59
	Test 03	✗	✓	10.60	13.30
	Test 04	✗	✗	45.80	52.40
	Average	-	-	17.43	23.17
BEATs+AASIST	Test 01	✓	✓	0.08	0.03
	Test 02	✓	✗	1.26	0.08
	Test 03	✗	✓	4.70	2.20
	Test 04	✗	✗	17.20	3.00
	Average	-	-	**5.81**	**1.33**

detection. Each system produces a score for a given audio clip, indicating the confidence that the clip is real. The EER is found by setting a threshold where the false acceptance rate equals the false rejection rate. Lower EERs indicate better detection performance of a system.

4.2. Performance of Baselines and Proposed System

We first focus on the performance of the two baselines AA-SIST and W2V2+AASIST in Table 4 under all test conditions. Although both models perform well under the seen condition of Test 01, their performance drops significantly under the unseen test conditions of Tests 02-04. It should also be noted that the introduction of W2V2 in AASIST generally improves the performance of speech deepfake detection as reported in [30], but the same trend is not observed for deepfake sound detection from the results observed in Table 4. This may be due to the fact that W2V2 was pre-trained on pure speech data, and speech exhibits significant differences in acoustic features compared to environmental sounds.

We now compare the performance of our proposed BEATs+AASIST system with the two baselines as reported in Table 4. It is evident that the introduction of BEATs significantly benefits the AASIST system, which was pre-trained on environmental sounds. This suggests that it is able to capture acoustic nuances more effectively than speech-based pre-trained models. BEATs+AASIST outperforms both baseline models across all test conditions, validating our hypothesis of leveraging an audio foundation model.

Let us consider having a closer look at the performances of all the systems under each test condition. The low EERs (below 1%) for all systems in Test 01 indicates that in-domain sound deepfake detection is a relatively simple task. In Test 02, the trained baselines have to handle fake data generated by unseen TTA and ATA models. The results show a significant drop in performance on Test 02 compared to Test 01, demonstrating the detection model's limited ability to generalize across different audio generation models. Similarly, in Test 03, the source datasets are out-of-domain for the trained detection models, leading to degraded performance compared to Test 01. Furthermore, we observe that AASIST and BEATs+AASIST systems are more affected by unseen source datasets than unseen
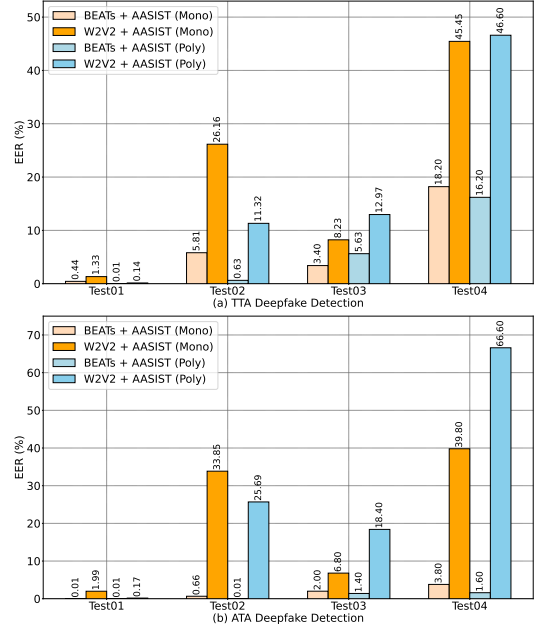

Figure 2: *Performance in EER (%) of W2V2+AASIST and BEATs+AASIST on monophonic and polyphonic data.*

audio generation models, whereas W2V2+AASIST shows an opposite trend indicating more dependence on generation models. Test 04 is the most challenging condition as both the audio generation models and the source datasets are out-of-domain, resulting significant drop in performance for all the systems. However, the proposed BEATs+AASIST system still performs substantially better than the two baselines.

4.3. Performance for Monophonic and Polyphonic Audio

In this work, the EnvSDD dataset contains two types of audio, monophonic and polyphonic, where the former has only a single sound event, while the latter exhibits multiple overlapping events. Figure 2 shows the detection performance of W2V2+AASIST and BEATs+AASIST on monophonic and polyphonic subsets. For BEATs+AASIST, the EERs on monophonic audio are generally higher than those on polyphonic audio, indicating that it is more challenging to detect monophonic deepfake audio. However, the EER on monophonic audio is lower than that on polyphonic audio for TTA deepfake detection in Test 03, where the source datasets are unseen during training. For W2V2+AASIST, the model performs better on polyphonic audio in Test 01 and Test 02, whereas the opposite trend is observed in Test 03 and Test 04, with better performance on monophonic audio. Overall, the proposed BEATs+AASIST outperforms W2V2+AASIST in all cases, further demonstrating the effectiveness of the pre-trained audio foundation model.

5. Conclusions

In this work, we have proposed EnvSDD, the first large-scale dataset with real and deepfake environmental audio clips for ESDD. We use two SOTA systems from the speech and singing voice deepfake detection domains as baselines and proposed a new system based on a pre-trained audio foundation model, BEATs. The results on EnvSDD under various test conditions demonstrate that the proposed system outperforms the two baselines. However, its generalization to unseen domains remains limited, which should be further explored in future work.

6. Acknowledgements

7. References

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

[2] H. Liu, R. Huang, Y. Liu *et al.*, "AudioLCM: Efficient and high-quality text-to-audio generation with minimal inference steps," in *Proc. ACM International Conference on Multimedia (ACM MM)*, 2024, pp. 7008–7017.

[3] J. Barnett, "The ethical implications of generative audio models: A systematic literature review," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 146–161.

[4] A.-T. Dao, M. Rouvier, and D. Matrouf, "ASVspoof 5 challenge: advanced ResNet architectures for robust voice spoofing detection," in *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof Workshop)*, 2024, pp. 163–169.

[5] J. Yamagishi, X. Wang, M. Todisco *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof Workshop)*, 2021, pp. 47–54.

[6] Y. Zhang, Y. Zang, J. Shi *et al.*, "SVDD 2024: The inaugural singing voice deepfake detection challenge," in *Proc. IEEE Spoken Language Technology Workshop (SLT Workshop)*, 2024, pp. 782–787.

[7] J. Yi, C. Wang, J. Tao *et al.*, "SceneFake: An initial dataset and benchmarks for scene fake audio detection," *Pattern Recognition*, vol. 152, p. 110468, 2024.

[8] H. Cheng, K. Li, L. Ye *et al.*, "EnvFake: An initial environmental-fake audio dataset for scene-consistency detection," in *Proc. IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 81–85.

[9] H. Ouajdi, O. Hadder, M. Tailleur *et al.*, "Detection of deepfake environmental audio," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2024, pp. 1–5.

[10] Z. Xie, B. Li, X. Xu *et al.*, "FakeSound: Deepfake general audio detection," *Proc. Interspeech*, 2024.

[11] C. D. Kim, B. Kim, H. Lee *et al.*, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[12] S. Chen, Y. Wu, C. Wang *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM International Conference on Multimedia (ACM MM)*, 2014, pp. 1041–1044.

[14] K. Choi, J. Im, L. Heller *et al.*, "Foley sound synthesis at the DCASE 2023 challenge," *arXiv preprint: 2304.12521*, 2023.

[15] T. Heittola, A. Mesaros, and T. Virtanen, "TAU Urban Acoustic Scenes 2019 Openset, Development dataset," *Zenodo*, 2019, doi:https://doi.org/10.5281/zenodo.2591503.

[16] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Sound Events 2017, Evaluation dataset," *Zenodo*, 2017, doi:https://doi.org/10.5281/zenodo.1040179.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Sound Events 2017, Development dataset," *Zenodo*, 2017, doi:https://doi.org/10.5281/zenodo.814831.

[19] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[20] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM International Conference on Multimedia (ACM MM)*, 2013, pp. 411–412.

[21] F. Kreuk, G. Synnaeve, A. Polyak *et al.*, "AudioGen: Textually guided audio generation," in *Proc. International Conference on Learning Representations (ICLR)*, 2023, pp. 1–16.

[22] H. Liu, Y. Yuan, X. Liu *et al.*, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[23] C.-Y. Hung, N. Majumder, Z. Kong *et al.*, "TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint: 2412.21037*, 2024.

[24] A. Q. Jiang, A. Sablayrolles, A. Mensch *et al.*, "Mistral 7B," *arXiv preprint:2310.06825*, 2023.

[25] Y. Wu, K. Chen, T. Zhang *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[26] P.-Y. Huang, H. Xu, J. Li *et al.*, "Masked autoencoders that listen," *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 35, pp. 28 708–28 720, 2022.

[27] H. Tak, J.-w. Jung, J. Patino *et al.*, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof Workshop)*, 2021, pp. 1–8.

[28] J.-W. Jung, H.-S. Heo, H. Tak *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 2022, pp. 6367–6371.

[29] X. Wang, H. Ji, C. Shi *et al.*, "Heterogeneous graph attention network," in *Proc. The World Wide Web Conference*, 2019, pp. 2022–2032.

[30] H. Tak, M. Todisco, X. Wang *et al.*, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2022, pp. 112–119.

[31] A. Babu, C. Wang, A. Tjandra *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 776–780.

[33] H. Tak, J. Patino, M. Todisco *et al.*, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

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