MultiAPI Spoof: A Multi-API Dataset and Local-Attention Network for Speech Anti-spoofing Detection

Xueping Zhang[1], *Zhenshan Zhang*[1], *Yechen Wang*[3], *Linxi Li*[3], *Liwei Jin*[3], *Ming Li*[1,2]

[1]Digital Innovation Research Center, Duke Kunshan University, China
[2]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3]OfSpectrum, Inc., USA

mingli369@cuhk.edu.cn

Abstract

Existing speech anti-spoofing benchmarks rely on a narrow set of public models, creating a substantial gap from real-world scenarios in which commercial systems employ diverse, often proprietary APIs. To address this issue, we introduce MultiAPI Spoof, a multi-API audio anti-spoofing dataset comprising about 230 hours of synthetic speech generated by 30 distinct APIs, including commercial services, open-source models, and online platforms. Furthermore, we propose Nes2Net-LA, a local-attention enhanced variant of Nes2Net that improves local context modeling and fine-grained spoofing feature extraction. Based on this dataset, we also define the API tracing task, enabling fine-grained attribution of spoofed audio to its generation source. Experiments show that Nes2Net-LA achieves state-of-the-art performance and offers superior robustness, particularly under diverse and unseen spoofing conditions. Code [1] and dataset [2] have been released.

Index Terms: Speech Anti-spoofing, Speech Deepfake Detection, MultiAPI Spoof, API Tracing, Local-Attention Network

1. Introduction

Recently, Text-To-Speech (TTS) [1, 2, 3, 4], Voice Conversion (VC) [5, 6, 7, 8], and generative modeling techniques [9, 10, 11, 12] have evolved rapidly. In particular, end-to-end dialogue systems [13, 14, 15, 16], speech continuation models [17, 18], and style- or emotion-specific speech generation models [19, 20, 21] have advanced significantly. As a result, synthetic speech has become increasingly realistic and pervasive in everyday applications. Modern audio generation systems, especially those based on diffusion and large-scale generative models [22, 14, 23], can now produce speech that closely mimics human prosody, timbre, and emotion. However, they have introduced serious security risks and can be easily misused for impersonation or misinformation.

Recent audio anti-spoofing approaches are typically based on a pre-trained model [24, 25, 26] that extracts high-level acoustic features. These features are then fed into a back-end classifier [27, 28, 29, 30] to distinguish bona fide from spoofed audio. Although recent studies in audio anti-spoofing have achieved notable progress, existing datasets [31, 32, 33, 34, 35, 36] are typically constructed from a limited number of public TTS or VC models, providing an incomplete view of today's real-world spoofing landscape. In practice, most industrial platforms adopt proprietary or closed-source APIs, making it difficult to access their model architectures, data pipelines, or synthesis mechanisms. Therefore, it remains unclear how well models trained on existing open-source benchmarking datasets will perform on real-world API data. Moreover, the rapid emergence of new generative paradigms results in a substantial domain gap between research benchmarks and real-world spoofing attacks.

To address these limitations, we introduce MultiAPI Spoof, a new multi-API speech anti-spoofing dataset designed for both anti-spoofing detection and API-level source tracing. Unlike prior datasets that focus on a few synthesis systems, MultiAPI Spoof comprises audio generated from 30 distinct APIs, including commercial TTS services, open-source speech models, and TTS websites. The dataset covers approximately 230 hours of spoofed speech. Based on the MultiAPI Spoof dataset, our contributions are as follows:

1. We show that there is a gap between previous research benchmarks and real-world spoofing scenarios; adding our API dataset in the training can also enhance the performance on current benchmarks.
2. Furthermore, we propose a new anti-spoofing detection method, namely Nes2Net-LA, built upon Nes2Net [30]. By integrating local attention modules between Nested blocks, Nes2Net-LA enhances local context modeling and fine-grained spoofing feature extraction, thereby improving robustness and discriminative capability. The Nes2Net-LA achieves state-of-the-art (SOTA) performance across multiple anti-spoofing benchmarks.
3. Finally, we introduce the API tracing task, which aims to identify the generation API of spoofed audio and establishes a benchmark for fine-grained source attribution.

2. MultiAPI Spoof Dataset

MultiAPI Spoof is a new multi-API audio anti-spoofing dataset designed to bridge the gap between research benchmarks and real-world synthetic speech. It contains approximately 230 hours of spoofed audio and an equal amount of bona fide speech from CommonVoice, maintaining a 1:1 balance between the two. All recordings are in English. The dataset provides a diverse set of spoofing conditions for both anti-spoofing detection and API-level source tracing.

2.1. Spoofed Audio Data Sources

The spoofed audio in MultiAPI Spoof is generated through 30 distinct APIs, reflecting a broad spectrum of synthesis techniques and real-world deployment scenarios:

1. Commercial TTS APIs: Speech synthesized by proprietary text-to-speech services widely used in industry.
2. Open-Source Models: Speech generated using publicly avail-

Corresponding Author: Ming Li

[1]https://github.com/XuepingZhang/MultiAPI-Spoof

[2]https://xuepingzhang.github.io/MultiAPI-Spoof-Dataset/

able neural TTS or voice conversion systems.

3. TTS Websites: Audio collected from online platforms providing web-based speech synthesis interfaces.

Each API corresponds to one labeled group (A0–A29), forming a comprehensive representation of modern TTS and generative pipelines.

2.2. Dataset Split

The MultiAPI Spoof dataset is partitioned by the APIs. APIs A0–A20 are used to construct the training, development, and evaluation subsets with a 70/10/20 % split, ensuring sufficient variation within seen sources. APIs A21–A23 are reserved entirely for development, while APIs A24–A29 are held out exclusively for evaluation. This design enables two evaluation conditions:

1. Seen evaluation, where systems are tested on spoofed samples generated from APIs that also appear in training (A0–A20).
2. Unseen evaluation, where systems are evaluated on spoofed samples from completely unseen APIs (A21–A29), allowing assessment of cross-source generalization.

3. Local Attention Enhanced Anti-spoofing Network

3.1. Prior Knowledge: Nested Res2Net (Nes2Net-X)

The Nes2Net-X architecture [30] is a multi-scale feature extractor for high-dimensional speech representations. An audio segment x_i is encoded into $x_i' \in \mathbb{R}^{C\times T'}$ and split into channel-wise subsets $x_{i,1}, \ldots, x_{i,J}$. Each subset is processed hierarchically: the first passes through Convolution(Conv) and Weighted Summation (WS), while subsequent subsets are fused with the previous output before convolution. All outputs are refined with a convolution and Squeeze-and-Excitation (SE) module with residual connections to obtain the anti-spoofing feature representations $h_{i,j} \in R^{(C/J)\times T'}$, as shown in (1).

$$\begin{aligned} z_{i,j} &= \mathrm{WS}\Big(\mathrm{Conv}\big(x_{i,j} + \mathbb{I}_{\{j>1\}}\, z_{i,j-1}\big)\Big), \\ h_{i,j} &= x_{i,j} + \mathrm{SE}\Big(\mathrm{Conv}(z_{i,j})\Big) \end{aligned} \tag{1}$$

3.2. Nes2Net with Local Attention (Nes2Net-LA)

While Nes2Net-X [30] effectively captures multi-scale structures, its refinement remains strictly hierarchical: each nested block only interacts with its immediate predecessor. This constrains long-range communication across blocks, which is increasingly important for high-dimensional speech representations. Hence, we propose to add Local Attention to the Nes2Net model (Nes2Net-LA).

As shown in Figure 1, for each block, we define a local sliding-window neighborhood $\mathcal{N}(i,j) = \{h_{i,k} \mid k \in [j-K, j+K]\}$, where K is the window radius. A local scaled dot-product self-attention (ATT) [38] operator is then applied to get the local feature representation $y_{i,j} \in \mathbb{R}^{(C/J)\times T'}$, as shown in (2).

$$y_{i,j} = \mathrm{ATT}\big(h_{i,j},\, \mathcal{N}(i,j)\big), \tag{2}$$

Finally, a residual connection aggregates the original representation $h_{i,j}$ and enhanced local feature representation $y_{i,j}$. All

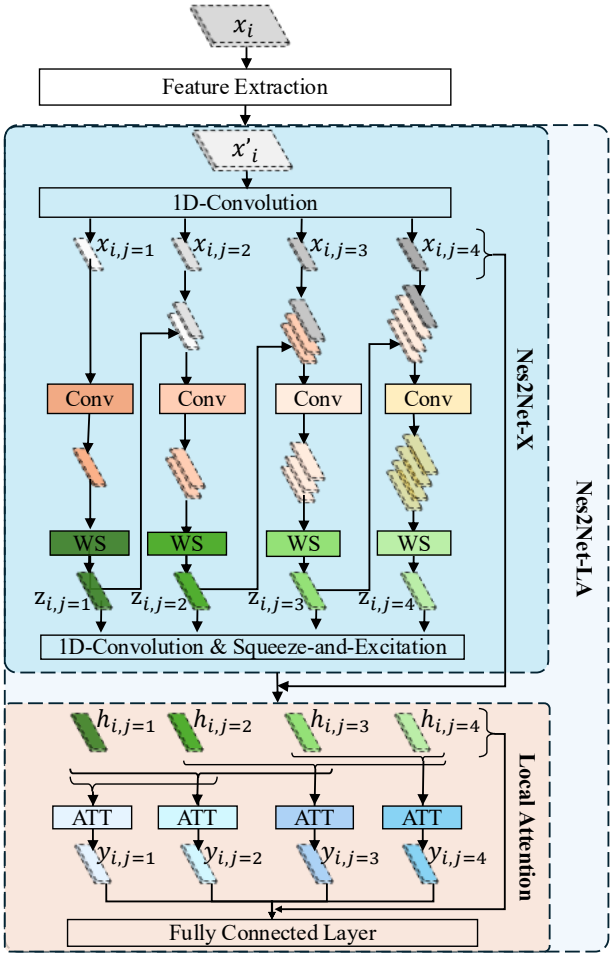

Figure 1: *Overall architecture of proposed Nes2Net-LA frameworks. The model first extracts high-dimensional representations from the input audio and then processes them using nested multi-scale feature fusion. Nes2Net-LA further enhances cross-block interactions through a sliding-window local attention mechanism. 'WS' represent Weighted Summation, and 'ATT' represent scaled dot-product self-attention.*

block outputs are concatenated and fed to a fully connected (FC) layer to produce the final anti-spoofing score.

Unlike global attention, which is too expensive for extended sequences of nested blocks, the proposed local attention considers only a small sliding window of neighboring blocks (e.g., 3). Within this window, each block can gather useful information from its nearby blocks and combine it with its own features. This approach makes the features more consistent and robust, improving the model's overall performance in anti-spoofing tasks.

4. Anti-Spoofing API Tracing Task

The anti-spoofing API tracing task aims to identify which API generated a given spoofed audio sample. Unlike conventional anti-spoofing, which only distinguishes bona fide and spoofed speech, API tracing provides fine-grained attribution. APIs are divided into seen and unseen sets. The seen set, consisting of 21 APIs (A0–A20), appears in training, while the unseen set is

Table 1: *Comparison of anti-spoofing performance without and with MultiAPI Spoof training set in training. Each cell in the table follows the format EER↓ / minDCF↓ / actDCF↓. The 'wo MultiAPI Spoof' setting trains models only on TIMIT, ODSS, FoR, AI4T, ASV5, and MLAAD, without any MultiAPI Spoof training set. The 'with MultiAPI Spoof' setting trains on the same data, plus the MultiAPI Spoof training set.*

Dataset	Model	ITW	MultiAPI Spoof			AI4T
			Seen	Unseen	Overall	
without MultiAPI Spoof	XLSR+AASIST [37]	2.02 / 0.026 / 0.029	-	-	7.30 / 0.098 / 0.106	12.96 / 0.132 / 0.190
	XLSR+Nes2Net [30]	1.73 / 0.023 / 0.025	-	-	7.08 / 0.098 / 0.103	7.77 / 0.093 / 0.110
	XLSR+Nes2Net-LA (Ours)	1.70 / 0.023 /0.020	-	-	6.11 / 0.085 / 0.089	7.76 / 0.090 / 0.099
with MultiAPI Spoof	XLSR+AASIST [37]	2.09 / 0.028 / 0.030	**0.48** / 0.007 / 0.0070	0.83 / 0.010 / 0.012	0.70 / 0.009 / 0.010	6.26 / 0.079 / 0.092
	XLSR+Nes2Net [30]	1.69 / 0.024 / 0.024	0.55 / 0.007 / 0.008	0.80 / 0.011 / 0.012	0.69 / 0.010 / 0.010	**5.64** / 0.052 / 0.079
	XLSR+Nes2Net-LA (Ours)	**1.42** / 0.020 / 0.021	**0.48** / 0.007 / 0.007	**0.62** / 0.009 / 0.009	**0.56** / 0.008 / 0.008	**5.64** / 0.051 / 0.077

reserved for evaluation to test generalization.

Our baseline model uses hidden representations from the XLSR-300M [24] encoder, followed by an attention pooling layer to aggregate embeddings from each encoder hidden layer, and ends with a Squeeze-and-Excitation (SE) layer to get the final results. During training, only the 21 seen APIs are used. At inference, samples whose maximum predicted probability falls below a threshold are classified as the unseen class, effectively turning the task into a 22-class classification problem.

5. Experiments

5.1. Experimental Setup

Dataset The anti-spoofing experiments are conducted on a collection of six public datasets: TIMIT [39], ODSS [40], FoR [41], AI4T [42], ASV5 [35], and MLAAD [43]. These corpora cover a wide variety of spoofing sources, including real-world collected data, text-to-speech (TTS), and voice conversion (VC). We consider two training configurations. In the first setting, the six datasets are merged into a single training set, and evaluation is performed across three target domains: the full ITW dataset [44], MultiAPI Spoof test set, and AI4T test set. In the second setting, the MultiAPI Spoof training set is additionally included in the training pool, and the models are re-evaluated on the same three domains.

For the API tracing task, both training and testing are performed on MultiAPI Spoof, and we report separate results for seen and unseen API categories to reflect generalization across API sources.

Processing All systems operate on normalized raw waveforms. Each audio sample is converted into a 4-second segment: signals shorter than 4 seconds are repeated until reaching the 4-second length, and longer signals are truncated. Unlike many existing audio anti-spoofing systems, we do not apply data augmentation in any of our experiments to ensure a clean, controlled comparison across models.

Training The anti-spoofing models evaluated in this work include XLSR+AASIST [37], XLSR+Nes2Net-X [30], and XLSR+Nes2Net-LA. All of those models have the same feature extroctor XLSR-300M [24]. For both Nes2Net-X and Nes2Net-LA, the number of channel splits is fixed at $J = 8$, and the local attention module uses a window size of $K = 1$. During training, XLSR+AASIST is optimized using Adam [45] with an initial learning rate of 1×10^{-6}, weight decay of 1×10^{-4}, and cross-entropy loss [46]. The XLSR+Nes2Net-X and XLSR+Nes2Net-LA systems use Adam with an initial learning rate of 5×10^{-6} and weight decay of 1×10^{-4}, and cross-entropy loss.

For the API tracing experiments, the model is trained using Adam with a learning rate of 1×10^{-5}, weight decay of 1×10^{-4}, and cross-entropy loss.

Metrics For the anti-spoofing task, performance is evaluated using Equal Error Rate (EER↓), minimum Decision Cost Function (minDCF↓), and actual Decision Cost Function (actDCF↓) [47]. For the API tracing task, we measure classification performance using precision, recall, and F1 [48]. Specifically, the F1 for seen APIs is computed as the macro-average of the F1 scores over the 21 seen API classes. For unseen APIs, F1 is computed on the single unseen class. The overall performance is reported as the macro-average across all classes, including the unseen class.

5.2. Experimental Results and Analysis

5.2.1. Anti-Spoofing on MultiAPI Spoof

To assess the value of MultiAPI Spoof for training and evaluation, we design two comparison experiments. The results are shown in Table 1.

In the first comparison experiment, models are trained on six commonly used anti-spoofing datasets (TIMIT [39], ODSS [40], FoR [41], AI4T [42], ASV5 [35], MLAAD [43]) without including MultiAPI Spoof training set. Two systems, XLSR+AASIST [37] and XLSR+Nes2Net-X [30], are evaluated on ITW [44], MultiAPI Spoof test set, and AI4T [42], respectively. Both models present relatively high EERs on the MultiAPI Spoof evaluation set, indicating a domain shift that existing datasets fail to cover. In the second comparison experiment, MultiAPI Spoof training set is incorporated into the training pool. Across all evaluation sets, especially on the MultiAPI Spoof test set itself, both XLSR+AASIST and XLSR+Nes2Net-X show substantial reductions in EER, minDCF, and actDCF. For instance, XLSR+AASIST decreases the EER from 7.30% to 0.70% on MultiAPI Spoof test set, and similarly, XLSR+Nes2Net decreases the EER from 7.08% to 0.69%.

Moreover, the benefits are not limited to MultiAPI Spoof. On the ITW dataset, XLSR+Nes2Net decreases from 1.73% to 1.69% in EER, and on AI4T, it decreases from 7.77% to 5.64%. Notably, within MultiAPI Spoof test set itself, the gains are observed not only on the seen sources but also on the unseen subset, indicating that the additional spoofing conditions contribute to more robust feature learning rather than overfitting to specific APIs. These consistent improvements across multiple evaluation sets suggest that adding MultiAPI Spoof in the train-

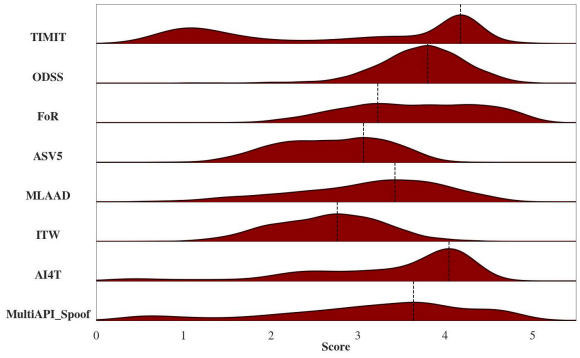

Figure 2: *Scoreq [49] distribution comparison across datasets. The dashed vertical line in each curve marks the peak density value.*

Table 2: *Comparison of our proposed XLSR+Nes2Net-LA system with recent state-of-the-art anti-spoofing models. "SP" denotes Sample Pruning [42]; "RB" denotes RawBoost augmentation [50]; "C" denotes codec augmentation. "Data Collection 1" consists of ASVspoof 2019 [33], FoR[41], ASVspoof 2021 DF[34], TIMIT[39], ODSS[40], MLAAD[43], and ASV5[35] training sets. On top of Data Collection 1, Data Collection 2 replaces ASVspoof 2019 and ASVspoof 2021 DF with high-quality AI4T [42] and the proposed MultiAPI Spoof training set.*

Model	Training data	Aug.	ITW	AI4T
XLSR+SLS [51]	ASVspoof 2019 LA	RB	7.46	N/A
XLSR+Mamba [29]	ASVspoof 2019 LA	RB	6.71	N/A
XLSR+AASIST [37]	ASVspoof 2019 LA	RB	10.46	N/A
XLSR+AASIST [37]	Data Collection 2	N/A	2.09	6.26
XLSR+LRC [42]	ASVspoof 2019	N/A	3.4	27.4
XLSR+LRC [42]	Data Collection 1	SP	1.70	12.4
XLSR+LRC [42]	Data Collection 1	SP & RB+C	1.90	10.2
XLSR+Nes2Net [30]	ASVspoof 2019	RB	5.52	N/A
XLSR+Nes2Net [30]	Data Collection 2	N/A	1.69	**5.64**
XLSR+ Nes2Net-LA (Ours)	Data Collection 2	N/A	**1.42**	**5.64**

ing effectively enhances cross-domain robustness and provides better generalization to unseen data.

To better understand this effect, we visualize the Scoreq[49] distributions of those datasets, as shown in Figure 2. MultiAPI Spoof exhibits a significantly broader quality distribution, spanning both low- and high-quality regions. Such diversity helps the model generalize better by preventing overfitting to narrow acoustic conditions, thereby improving detection performance in more realistic, heterogeneous environments.

5.2.2. *Effectiveness of Local Attention (Nes2Net-LA)*

As shown in Table 1 and Table 2, Nes2Net-LA trained on the data collection pool outperforms recent state-of-the-art models across all evaluation benchmarks, even without any data augmentation or pruning. The most substantial improvements are observed on the unseen split of the MultiAPI Spoof test set. These results demonstrate that the proposed local attention mechanism produces more discriminative and robust anti-spoofing representations.

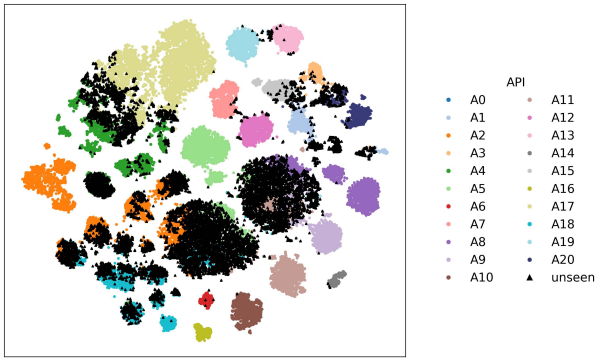

Figure 3: *t-SNE [52] visualization of XLSR-extracted embeddings for the MultiAPI Spoof eval set. Unseen APIs are A24-A29*

Table 3: *API Tracing Performance on the MultiAPI Spoof Dataset. Seen APIs correspond to A0–A20, while dev unseen APIs are A21–A23, and eval unseen APIs are A24–A29.*

	dev			eval		
	precision↑	**recall↑**	**F1↑**	**precision↑**	**recall↑**	**F1↑**
seen	0.950	0.924	0.937	0.950	0.923	0.936
unseen	**0.959**	0.467	0.628	**0.972**	0.520	0.678
overall	0.778	0.910	0.785	0.770	0.917	0.782

5.2.3. *API Tracing on MultiAPI Spoof*

Table 3 presents the API tracing results on MultiAPI Spoof. The model is trained on the MultiAPI Spoof training set and tested on the dev and eval sets. We report results separately for seen and unseen API types. Overall performance is high across seen APIs, with both dev and eval achieving substantial precision, recall, and F1 scores. However, high precision but low recall for the unseen class shows that predictions are accurate, but many unseen-class instances are not correctly identified and are falsely rejected as unseen cases. This indicates that current methods for this task, particularly in handling unseen APIs, still require further investigation.

As shown in Figure 3, t-SNE [52] visualizations further reveal that embeddings of unseen APIs do not form separable clusters; instead, they are mixed with multiple seen categories. This suggests that the model primarily learns API-specific acoustic cues and struggles to generalize to unseen APIs whose acoustic or behavioral signatures differ significantly from the training distribution. These findings highlight the challenge of zero-shot API tracing and suggest that future models require stronger invariant representation learning.

6. Conclusion

In this paper, we present MultiAPI Spoof, a multi-API speech anti-spoofing dataset, and further introduce the API tracing task for fine-grained source attribution. Experiments show that incorporating MultiAPI Spoof into training significantly improves cross-domain robustness. We also propose a local-attention enhanced anti-spoofing network, namely Nes2Net-LA. It outperforms Nes2Net-X, achieving state-of-the-art performance, demonstrating its effectiveness in improving robustness and discriminative capability.

7. Generative AI Use Disclosure

Large LanguageModels (LLMs) were used solely for manuscript polishing (e.g., rephrasing and grammar checks) to improve clarity and readability. The LLMs were not used for ideation, methodology, experimental design, data analysis, or result interpretation. All scientific content was produced and verified by the authors.

8. Acknowledgments

Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

9. References

[1] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo *et al.*, "Qwen3-tts technical report," *arXiv preprint arXiv:2601.15621*, 2026.

[2] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2025, pp. 6255–6271.

[3] H. Li, C. Jin, C. Li, W. Guan, Z. Huang, and X. Chen, "Restyle-tts: Relative and continuous style control for zero-shot speech synthesis," *arXiv preprint arXiv:2601.03632*, 2026.

[4] Z. Liu, S. Wang, P. Zhu, M. Bi, and H. Li, "E1 tts: Simple and fast non-autoregressive tts," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[5] J. Yao, Y. Yuguang, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, "Stablevc: Style controllable zero-shot voice conversion with conditional flow matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25 669–25 677.

[6] J. Kim, J.-H. Kim, Y. Choi, T. D. Nguyen, S. Mun, and J. S. Chung, "Adaptvc: High quality voice conversion with adaptive learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[7] Z. Wang, T. Li, W. Ge, Z. Cui, S. Zhang, and J. Feng, "Onevoice: One model, triple scenarios-towards unified zero-shot voice conversion," *arXiv preprint arXiv:2601.18094*, 2026.

[8] J. Yao, Y. Yuguang, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, "Stablevc: Style controllable zero-shot voice conversion with conditional flow matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25 669–25 677.

[9] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[10] R. Cai, Y. Lin, Y. Wang, C. Fu, and X. Zeng, "Unifying speech recognition, synthesis and conversion with autoregressive transformers," 2026. [Online]. Available: https://arxiv.org/abs/2601.10770

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[13] T. F. Team, Q. Chen, L. Cheng, C. Deng, X. Li, J. Liu, C.-H. Tan, W. Wang, J. Xu, J. Ye *et al.*, "Fun-audio-chat technical report," *arXiv preprint arXiv:2512.20156*, 2025.

[14] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[15] Y. Li, S. Ji, Y. Chen, T. Liang, H. Ying, Y. Wang, J. Li, J. Fang, and Z. Zhao, "Wavbench: Benchmarking reasoning, colloquialism, and paralinguistics for end-to-end spoken dialogue models," *arXiv preprint arXiv:2602.12135*, 2026.

[16] Y. Li, J. Liu, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, L. Ming, G. Dong, D. Pan *et al.*, "Baichuan-omni-1.5 technical report," *arXiv preprint arXiv:2501.15368*, 2025.

[17] T. Li, J. Liu, T. Zhang, Y. Fang, D. Pan, M. Wang, Z. Liang, Z. Li, M. Lin, G. Dong *et al.*, "Baichuan-audio: A unified framework for end-to-end speech interaction," *arXiv preprint arXiv:2502.17239*, 2025.

[18] L.-C.-T. Xiaomi, "Mimo-audio: Audio language models are few-shot learners," 2025. [Online]. Available: https://github.com/XiaomiMiMo/MiMo-Audio

[19] H. Xie, H. Lin, W. Cao, D. Guo, W. Tian, J. Wu, H. Wen, R. Shang, H. Liu, Z. Jiang, Y. Jiang, W. Chen, R. Yan, J. Qian, Y. Yan, S. Yin, M. Tao, X. Chen, L. Xie, and X. Wang, "Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity," *arXiv preprint arXiv:2510.23541*, 2025.

[20] D. Chen, X. Zhang, Y. Wang, K. Dai, L. Ma, and Z. Wu, "Flexivoice: Enabling flexible style control in zero-shot tts with natural language instructions," *arXiv preprint arXiv:2601.04656*, 2026.

[21] C. Yan, B. Wu, P. Yang, P. Tan, G. Hu, Y. Zhang, Xiangyu, Zhang, F. Tian, X. Yang, X. Zhang, D. Jiang, and G. Yu, "Step-audio-editx technical report," *arXiv preprint arXiv:2511.03601*, 2025.

[22] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generative models through direct preference optimization," in *ACM Multimedia*, 2024.

[23] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[26] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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.

[27] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[28] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2022, pp. 6367–6371.

[29] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*.

[30] T. Liu, D.-T. Truong, R. K. Das, K. A. Lee, and H. Li, "Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 12 005–12 018, 2025.

[31] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2022.

[32] Z. Li, Y. Lin, T. Yao, H. Suo, P. Zhang, Y. Ren, Z. Cai, H. Nishizaki, and M. Li, "The database and benchmark for the source speaker tracing challenge 2024," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 1254–1261.

[33] 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 Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[34] 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.

[35] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen *et al.*, "Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *Proceedings of ASVspoof*, 2024, pp. 1–8.

[36] H. Wu, Y. Tseng, and H.-y. Lee, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," in *Proceedings of Interspeech*, 2024, pp. 1770–1774.

[37] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. W. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proceedings of Odyssey*, 2022.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[39] 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, pp. 50 851–50 866, 2023.

[40] A. Yaroshchuk, C. Papastergiopoulos, L. Cuccovillo, P. Aichroth, K. Votis, and D. Tzovaras, "An open dataset of synthetic speech," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023, pp. 1–6.

[41] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–10.

[42] D. Combei, A. Stan, D. Oneata, N. Müller, and H. Cucu, "Unmasking real-world audio deepfakes: A data-centric approach," in *Proceedings of Interspeech*, 2025, pp. 5343–5347.

[43] 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 *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–7.

[44] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *Proceedings of Interspeech*, pp. 2783–2787, 2022.

[45] D. Kingma, "Adam: a method for stochastic optimization," in *International Conference Learn Represent*, 2014.

[46] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*, 2023, pp. 23 803–23 828.

[47] H. Delgado, N. Evans, J.-w. Jung, T. Kinnunen, I. Kukanov, K. A. Lee, X. Liu, H.-j. Shim, M. Sahidullah, H. Tak *et al.*, "Asvspoof 5 evaluation plan," https://www.asvspoof.org/file/ASVspoof5___Evaluation_Plan_Phase2.pdf, 2024, [Online].

[48] P. Christen, D. J. Hand, and N. Kirielle, "A review of the f-measure: its history, properties, criticism, and alternatives," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–24, 2023.

[49] A. Ragano, J. Skoglund, and A. Hines, "Scoreq: Speech quality assessment with contrastive regression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105 702–105 729, 2024.

[50] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6382–6386.

[51] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised xls-r and sls classifier," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6765–6773.

[52] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.