STOPA: A Dataset of Systematic VariaTion Of DeePfake Audio for Open-Set Source Tracing and Attribution

Anton Firc[1], *Manasi Chhibber*[2], *Jagabandhu Mishra*[2], *Vishwanath Pratap Singh*[2], *Tomi H. Kinnunen*[2], *Kamil Malinka*[1]

[1]Brno University of Technology, Czech Republic
[2]University of Eastern Finland, Finland

ifirc@fit.vut.cz, manasi.chhibber@uef.fi, jagabandhu.mishra@uef.fi, vsingh@uef.fi, tomi.kinnunen@uef.fi, malinka@fit.vut.cz

Abstract

A key research area in deepfake speech detection is source tracing — determining the origin of synthesised utterances. The approaches may involve identifying the acoustic model (AM), vocoder model (VM), or other generation-specific parameters. However, progress is limited by the lack of a dedicated, systematically curated dataset. To address this, we introduce STOPA, a systematically varied and metadata-rich dataset for deepfake speech source tracing, covering 8 AMs, 6 VMs, and diverse parameter settings across 700k samples from 13 distinct synthesisers. Unlike existing datasets, which often feature limited variation or sparse metadata, STOPA provides a systematically controlled framework covering a broader range of generative factors, such as the choice of the vocoder model, acoustic model, or pretrained weights, ensuring higher attribution reliability. This control improves attribution accuracy, aiding forensic analysis, deepfake detection, and generative model transparency.

Index Terms: source tracing, dataset, anti-spoofing, synthetic speech, deepfake

1. Introduction

The proliferation of deepfakes has led to increasingly realistic synthetic speech, posing significant threats to humans and speaker recognition. Malicious actors exploit deepfake speech for impersonation or fraud, with AI-generated voices even used to initiate fraudulent financial transactions [1].

Various detection approaches, from handcrafted feature-based to deep learning, have been developed to counter these threats [2, 3]. While substantial progress has been made, they provide limited insight into *how* or *who* generated given speech. Unlike deepfake detection, which determines whether an utterance is real or fake, **source tracing** (or attack attribution) assumes this classification has already been made. The aim is to identify the entity or system responsible for generating the speech. These methods thus offer deeper insights into the generation process, thereby enhancing forensic investigations [4, 5].

Despite recent progress, the lack of specialized datasets limits source tracing research. Existing datasets (reviewed in Section 2.2) focus on detection but lack **systematic variation** in acoustic models, vocoders, and other generative parameters. For example, the widely-adopted ASVspoof dataset series [2, 3, 6] is collected through semi-controlled crowdsourcing, where attack crafting (i.e. text-to-speech or voice conversion system development) follows a common generation protocol—but with lot of flexibility methods and their implementations, chosen by the data contributors. Although this design suits deepfake detection tasks, it prevents systematic analysis of attack characteristics essential for source tracing.

To address this research gap, we introduce **STOPA**[1] , a new dataset specifically designed for source tracing in deepfake audio detection. Our dataset features a systematic variation of key generative components, allowing researchers to analyze the impact of different model choices. For example, we include four distinct attacks generated using Tacotron2 [7]:

- Pairs of attacks share the same Tacotron2 implementation but differ in their vocoder choice.
- Tacotron2 models vary in architecture parameters (e.g., the number of attention heads) and pretrained weights.

While much of the recent focus has been on newer text-to-speech (TTS) models, legacy systems like Tacotron2 remain relevant due to their stable implementations, widespread availability, and ability to produce high-quality speech [8]. If detection systems fail to handle these well-established models, they present an attractive option for attackers.

A key difference of STOPA compared to existing datasets and most of the published methodology literature relies on the evaluation protocols design, intended to promote development of source tracing methods for 'open-world' settings. To elaborate, the majority of the source-tracing literature approaches the task as a multiclass task [4, 5, 9, 10, 11] assuming known attacks during training and handling unknown attacks via an additional fallback class with a dedicated out-of-distribution classifier. Whereas this treatment leads to class priors dependencies in performance metrics, extra classifiers further complicate system design and evaluation.

Inspired by the established principles of the speaker detection evaluation benchmarks coordinated by NIST [12], we frame source tracing as an **open-world detection task** that unifies source tracing and 'detecting the unknown' under a simple and coherent evaluation framework. A test instance is evaluated against a hypothesized spoofing attack, vocoder, or acoustic model. The task is to answer whether or not the test instance originates from the hypothesized source generator (null hypothesis) or not (alternative hypothesis). Similar to NIST SREs, designed to support 'continual growing' of a speaker database, our protocol forbids the use of *other* attacks for training or testing purposes: a decision must solely be made by comparing the hypothesized model against a single test instance. This mindset intends to promote the design of deepfake profiling approaches where the database of deepfake 'signatures' (similar to databases of computer viruses or malware) is allowed to grow dynamically during the lifespan of the profiling system. A 'fixed' dataset of deepfake generators, addressed as a multiclass task, requires re-training the profiling system whenever a new deepfake generator is encountered.

STOPA's main contributions include: *(1)* a scalable eval-

[1]https://doi.org/10.5281/zenodo.15462919

Table 1: *Overview of related datasets. Column Var. indicates whether the dataset systematically varies acoustic models (AMs) and vocoders (VMs) to facilitate source tracing.*

Dataset	#Utt.	#Sys.	#Spk.	#AM	#VM	Var.	Avail.
ASVspoof19 LA [2]	121K	19	107	19	11	✗	Pub.
ASVspoof21 DF [6]	593K	100+	93	19+	11+	✗	Pub.
ASVspoof5 [3]	1.2M	32	1,922	22	N/A	✗	Pub.
SemaFor [13]	17K	11	25	11	N/A	✗	Pub.
MLAAD [14]	154K	82	N/A	13	6	✗	Pub.
TIMIT-TTS [15]	79K	12	37	12	2	✗	Pub.
[16]	1.8K	5+	9	N/A	N/A	✗	Int.
[5]	63K	8	692	1	8	✗	Int.
[17]	N/A	17	N/A	3	14	✓	Int.
STOPA (ours)	**699K**	**13**	**107**	**8**	**6**	✓	**Pub.**

uation protocol enabling open-world source tracing by integrating new attack signatures without retraining, *(2)* systematic variation in spoofed speech with diverse acoustic models, vocoders, and hyperparameters, and *(3)* extensive metadata for fine-grained analysis and benchmarking of synthetic speech.

The benchmark results and code base are available at `https://github.com/Manasi2001/STOPA`.

2. Related work

2.1. Speech synthesis and voice conversion

Speech synthesis typically follows a two-stage pipeline: an acoustic model (AM) generates a spectrogram, which a vocoder model (VM) converts into an audible waveform [18]. AMs fall into two main categories: text-to-speech (TTS) and voice conversion (VC) [1]. TTS models transform the text into a speech representation. Relevant AMs include autoregressive models (e.g., Tacotron2 [7]), non-autoregressive models (e.g., FastPitch [19]), flow-based models (e.g., GlowTTS [20, 21]), and diffusion-based models (e.g., DiffGAN-TTS [22]).

In contrast, VC algorithms generate speech by combining linguistic content from one input with the vocal characteristics of another. These models may use representation learning techniques (e.g., SpeechSplit [23]) or diffusion-based methods (e.g., Diff-VC [24]).

Most commonly used vocoders include rule-based approaches (e.g., Griffin-Lim [25]), GAN-based methods (e.g., MelGAN [26], HiFi-GAN [27]), and diffusion-based models (e.g., WaveGrad [28]).

2.2. Deepfake audio datasets

The most widely recognized and adapted datasets for anti-spoofing are the ASVspoof datasets [2, 3, 6]. Additionally, MLAAD [14], TIMIT-TTS [15], and SemaFor [13] provide additional synthetic speech collections. A comparison with related datasets is shown in Table 1. While effective for anti-spoofing, these datasets were not designed for source tracing. However, some have been repurposed for this task.

The Interspeech 2025 special session on source tracing[2] uses MLAAD [14] dataset as a benchmark, incorporating a source tracing protocol. MLAAD includes 38 languages and samples generated from 82 different synthesis systems spanning 38 architectures, ranging from traditional models like Tacotron2 and Griffin-Lim to state-of-the-art approaches such as XTTS and VITS. The dataset introduces variation through the language dimension, enabling analysis of synthesis methods across

[2] `https://deepfake-total.com/sourcetracing`

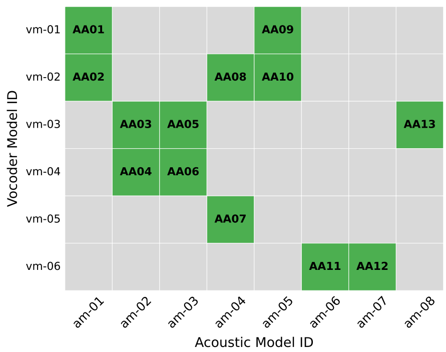

Figure 1: *Combination of AMs and VMs for attacks*

Table 2: *Used model architectures and implementations*

ID	Type	Architecture	Implementation
am-01	TTS	Tacotron2 [7]	[31]
am-02	TTS	Tacotron2 [7]	[32]
am-03	TTS	DiffGAN-TTS[22]	[33]
am-04	TTS	GlowTTS [20]	[31]
am-05	TTS	FastPitch [19]	[31]
am-06	TTS	GlowTTS [21]	[34]
am-07	VC	SpeechSplit [23]	[35]
am-08	VC	Diff-VC [24]	[36]
vm-01	VM	WaveGrad [28]	[31]
vm-02	VM	GriffinLim [25]	[31]
vm-03	VM	HiFi-GAN [27]	[32]
vm-04	VM	MelGAN [26]	[37]
vm-05	VM	HiFi-GAN [27]	[31]
vm-06	VM	ParallelWaveGAN [38]	[39]

different training corpora and model configurations. The baseline system [29] reports a 63% EER, providing a reference for source tracing performance on large-scale multilingual datasets.

Additionally, several studies address source tracing using existing datasets. ASVspoof19 LA has been utilised in [4, 9, 10], SemaFor in [10, 11], MLAAD in [30], and other researchers [5, 16, 17] created and used internal datasets. STOPA thus complements existing datasets by introducing systematic variation in acoustic models, vocoders, and hyperparameters, enabling fine-grained analysis. Unlike prior works, it employs an open-world evaluation protocol, as a step towards solutions expected in audio forensics and other applications.

3. STOPA dataset description

3.1. Creation procedure

We introduce a new dataset for synthetic speech source tracing, systematically varying acoustic and vocoder models across multiple two-stage speech synthesis tools (Figure 1, Table 2). Unlike end-to-end models, which jointly optimise all synthesis components, two-stage pipelines separate the acoustic model and vocoder, making it possible to analyze how each stage contributes to synthesis variations.

To model real-world synthetic speech generation while maintaining a systematic variation, we use pretrained models from the VCTK dataset [40]. Rather than retraining all models under identical conditions, leveraging existing models ensures representative variability of real-world synthesisers while maintaining strict control over evaluation protocols. We focus on widely used synthesis pipelines, selecting combinations of

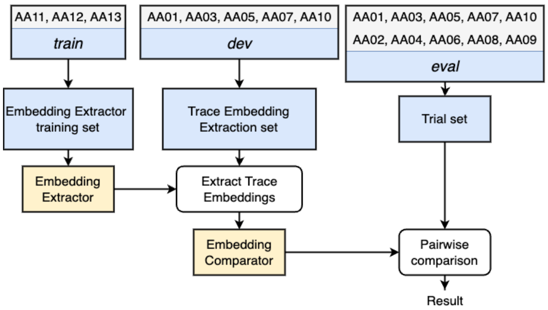

Figure 2: *Dataset structure visualisation.*

acoustic and vocoder models that are commonly available online. This approach reflects possible real-world attackers, ensuring that evaluations remain practically relevant.

We generate speech using 13 tools, each representing a distinct attack—a unique combination or parametrization[3] of acoustic and vocoder models. The dataset follows the VCTK corpus [40] and adopts the ASVspoof2019 Logical Access (LA) data generation protocol [2]. STOPA is attack-disjoint from ASVspoof2019 LA, ensuring no direct overlap in synthetic speech generation. While speaker identities overlap between ASVspoof2019's *dev* set and our trace embedding extraction set (Section 3.3), this should not introduce data leakage, as speaker identity is not a discriminative feature in source tracing. Consequently, source tracing models pretrained on ASVspoof2019 remain compatible with our dataset, provided that attribution focuses on synthesis methods rather than speaker-dependent characteristics.

We collected pretrained model weights from publicly available implementations on GitHub, using only those trained on the VCTK dataset. We then followed the ASVspoof2019 data generation protocols to create seen (*trn*, *dev*) and unseen (*eval*) subsets. Utterances not included in the pretrained weights or missing from the VCTK dataset were excluded, as detailed in the dataset README. For synthesisers requiring a target speaker utterance (AA11, AA13), we selected the longest available bonafide utterance for that speaker. Since we rely on pretrained weights, we had no control over the original training speakers, making speaker leakage between the *train*, *dev*, and *eval* subsets likely.

To manage dataset size while preserving essential variability, we applied a selection process. As part of this process, we included common utterances, having identical linguistic content across TTS and VC attacks, while selecting the remaining utterances at random. Incorporating common utterances enables content-controlled analysis and studies on the impact of identical linguistic content across different attacks. This approach is useful for examining attack consistency, evaluating content-dependent spoofing traits, and, in our case, investigating whether shared utterances enhance attribution performance. The number of common utterances was determined using the elbow criterion on pairwise overlap values, balancing their presence while avoiding excessive overlap between attacks.

3.2. Post-processing

We applied a post-processing pipeline to detect and mitigate shortcut artifacts [3], analyzing the following features: silence durations, peak amplitude, utterance duration, and energy.

[3]Different parameters (e.g., number of attention heads) or weights.

Table 3: *Count of trials by evaluation conditions. TE denotes a trace embedding.*

Condition	**Trials**	**Condition**	**Trials**
Same attack	3,149,000	Different attack	28,341,000
Same AM	6,298,000	Different AM	25,192,000
Same VM	6,298,000	Different VM	25,192,000
Known TE	15,745,000	Unknown TE	15,745,000
All trials		31,490,000	

For each waveform $\mathbf{x}$, we removed DC bias and normalized amplitude before detecting non-speech segments. Speech onset s^* and offset e^* were determined based on voice activity detection (VAD) activation or root mean square energy (RMSE) exceeding a threshold θ_{rmse}. Leading and trailing silence were identified by counting consecutive non-speech frames, with stable RMSE over W frames signalling the end of the speech. Empirical thresholds ($\theta_{\text{rmse}} = 0.05, W = 18$) were optimized to remove non-speech while preserving intelligibility. We evaluated intelligibility using Word Error Rate (WER) with Whisper-large-v3 [41] and JiWER [42]. Samples with WER > 100 (1%) were removed from *train*/*dev* sets while preserving speaker balance but retained in *eval*, as attribution models should also be able to handle degraded inputs.

3.3. Detection protocol and evaluation conditions

The dataset is organized to support source tracing as a detection task. Its structure follows the original *train*, *dev*, and *eval* subsets, as shown in Figure 2. The evaluation protocol defines conditions under which comparisons are made, specifying training, reference, evaluation partitions, and the creation of trial pairs.

The training set consists of three attacks and 20 speakers not present in any further sets. The reference set, derived from the original *dev* set, includes varying conditions to evaluate different embedding extraction methods. To systematically assess these methods, trace embedding models vary across two dimensions: (1) *the amount of data used*, ranging from a single utterance per attack to all available utterances, and (2) *the utterance selection strategy*, which either uses common utterances across all attacks or a more realistic scenario where utterances are randomly sampled for each attack.

Finally, the trial set, derived from the original *eval* subset, defines pairwise comparison for assessing source tracing performance. It includes 10 attack types: five known (present in the reference set) and five unknown. The protocol assigns multiple labels (same attack, same AM, or same VM), enabling analysis at different levels of granularity. The protocol defines 31,490,000 trials, with detailed statistics provided in Table 3.

3.4. Dataset compatibility with pretrained SSL models

Self-supervised learning (SSL) models are widely used in deepfake detection, but training data overlap with STOPA must be avoided to prevent evaluation bias. Several SSLs, including XLSR-Wav2Vec2 [43], WavLM [44], and Whisper [41], have no exposure to VCTK, ensuring broad compatibility.

4. Demonstration of STOPA Use

While our core contribution is the design and construction of STOPA itself, it is necessary to demonstrate its use in the open-world evaluation setting. Our purpose is *not* to optimize any

Table 4: *Pooled EER (%) for used systems across all conditions. AASIST CM denotes the ASVspoof2019-trained countermeasure system.*

	System	ATK	AM	VM
Known Attacks	ResNet-34	47.61	47.61	47.81
	AASIST CM	39.15	39.15	39.68
	AASIST STOPA	47.05	47.05	45.25
Unknown Attacks	ResNet-34	49.55	50.32	49.14
	AASIST CM	35.34	38.69	37.31
	AASIST STOPA	47.75	49.37	48.68

systems for best performance but to provide pilot results and highlight novel challenges brought in by STOPA.

We address attack type, acoustic model, and vocoder model classification using three 'trace embedding' extractors: AASIST CM [45] countermeasure (trained on ASVspoof2019 LA for binary spoof detection), and ResNet-34 [46] and AASIST STOPA, both trained on the STOPA training partition for attribution. ResNet-34 and AASIST STOPA are trained for 100 epochs. We emphasize that none of these models have been exposed to the attacks used in the source tracing reference and trial sets, as the STOPA dataset is explicitly designed to be attack-disjoint. The 'zero-shot' mindset (the ability to add new attacks without needing to retrain the embedding extractor) is a key design factor of STOPA protocols.

The actual trace embeddings for each attack are extracted using the predefined models. Following the simplest known classifier from speaker detection literature, we form a model for each attack by averaging the training embeddings. Test utterances are then compared to each reference using cosine similarity. A *positive* trial is where the ground-truth label (attack, vocoder, or acoustic model ID) in the model and the test match; otherwise, it is a *negative* trial. The negative set is further split into two subsets: (1) a different known attack and (2) an unknown attack, allowing for EER computation in both scenarios. The latter represents 'the unknown unknown'.

Table 4 shows error rates for detecting attack type (ATK), acoustic model (AM), and vocoder model (VM). The very high EERs reflects the difficulty of our open-world task setting. The best performance was obtained with the ASVspoof2019-trained AASIST CM. Both AASIST STOPA and the ResNet-34 models solely trained on STOPA produce, essentially, a chance rate. Since ASVspoof 2019 contains a larger number of training attacks, this strongly suggests that the more limited set of 3 STOPA attacks is not sufficient for capturing 'general variability' of attack-specific traits. Furthermore, using common vs. non-common trace embedding utterances has no impact on final performance.

A further t-SNE analysis (Figure 3) reveals a complex embedding space for ASVspoof2019-trained AASIST CM, where only AA10 forms a clear cluster. Training on STOPA did not form any clusters, leading to increased EER. Thus, the mean embedding approach was ineffective due to this extensive cluster overlap.

It might be useful to reflect these results on those typically obtained from speaker detection studies. Modern speaker embedding extractors, such as ECAPA-TDNN, are typically trained using *thousands* of speakers (classes). This leads to a discriminative 'speaker trait' space that transfers easily to classify previously unseen speakers. In contrast, we used only a handful of attacks (3 in STOPA) in training.

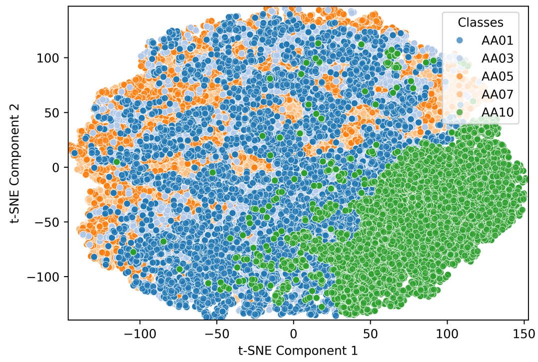

Figure 3: *t-SNE trace embedding space visualisation for the ASVspoof2019-pretrained AASIST model.*

5. Conclusions

We introduced STOPA, a novel dataset for synthetic speech source tracing, featuring 13 distinct attacks with systematic variations in acoustic and vocoder models. Unlike existing datasets, STOPA is designed to address 'open-world' source tracing. Inspired by NIST speaker detection benchmarks, our evaluation protocols unify source tracing and unknown attack detection under a scalable framework. By prohibiting the use of other attack data for training or testing, we encourage the development of dynamic 'attack signature databases' that evolve over time, similar to antivirus systems.

Our evaluation shows that established models like AASIST and ResNet fail to structure the signature space, resulting in attack detection EERs exceeding 30%. This demonstrates the fundamental challenge of zero-shot source tracing and highlights the need for improved embedding extraction and comparison strategies. STOPA thus provides a critical benchmark for advancing scalable and adaptive source tracing systems.

6. Acknowledgements

This work was partially supported by the Brno University of Technology (internal project FIT-S-23-8151) and the Academy of Finland (DecisionNo. 349605, project "SPEECHFAKES"). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

7. References

[1] A. Firc, K. Malinka, and P. Hanáček, "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors," *Heliyon*, vol. 9, no. 4, p. e15090, 2023.

[2] X. Wang, J. Yamagishi, M. Todisco *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[3] X. Wang, H. Delgado, H. Tak *et al.*, "Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *arXiv preprint*, vol. arXiv:2502.08857, 2025.

[4] M. Chhibber, J. Mishra, H. Shim *et al.*, "An explainable probabilistic attribute embedding approach for spoofed speech characterization," *arXiv preprint*, vol. arXiv:2409.11027, 2024.

[5] X. Yan, J. Yi, J. Tao *et al.*, "An initial investigation for detecting vocoder fingerprints of fake audio," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, ser. DDAM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 61–68.

[6] H. Delgado, N. Evans, T. Kinnunen *et al.*, "Asvspoof 2021 challenge - speech deepfake database," 2021.

[7] J. Shen, R. Pang, R. J. Weiss *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[8] A. González-Docasal, A. Álvarez, and H. Arzelus, "Exploring the limits of neural voice cloning: A case study on two well-known personalities," in *IberSPEECH 2022*, 2022, pp. 11–15.

[9] C. Borrelli, P. Bestagini, F. Antonacci *et al.*, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, no. 1, p. 2, Apr 2021.

[10] K. Bhagtani, E. R. Bartusiak, A. K. S. Yadav *et al.*, "Synthesized speech attribution using the patchout spectrogram attribution transformer," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*. New York, NY, USA: Association for Computing Machinery, 2023, p. 157–162.

[11] E. R. Bartusiak and E. J. Delp, "Transformer-based speech synthesizer attribution in an open set scenario," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 329–336.

[12] C. S. Greenberg, L. P. Mason, S. O. Sadjadi *et al.*, "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Computer Speech & Language*, vol. 60, p. 101032, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302761

[13] W. Corvey, "Semafor," Aug 2020. [Online]. Available: https://www.darpa.mil/research/programs/semantic-forensics

[14] N. M. Müller, P. Kawa, W. H. Choong *et al.*, "Mlaad: The multi-language audio anti-spoofing dataset," *International Joint Conference on Neural Networks (IJCNN)*, 2024.

[15] D. Salvi, B. Hosler, P. Bestagini *et al.*, "Timit-tts: A text-to-speech dataset for multimodal synthetic media detection," *IEEE Access*, vol. 11, pp. 50 851–50 866, 2023.

[16] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[17] C. Y. Zhang, J. Yi, J. Tao *et al.*, "Distinguishing neural speech synthesis models through fingerprints in speech waveforms," *arXiv preprint*, vol. arXiv:2309.06780, 2024.

[18] A. Firc, K. Malinka, and P. Hanáček, "Diffuse or confuse: A diffusion deepfake speech dataset," in *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2024, pp. 1–7.

[19] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[20] E. Casanova, C. Shulby, E. Gölge *et al.*, "Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model," in *Interspeech 2021*, 2021, pp. 3645–3649.

[21] J. Kim, S. Kim, J. Kong *et al.*, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8067–8077.

[22] S. Liu, D. Su, and D. Yu, "Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans," *arXiv preprint*, vol. arXiv:2201.11972, 2022.

[23] K. Qian, Y. Zhang, S. Chang *et al.*, "Unsupervised speech decomposition via triple information bottleneck," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 7836–7846.

[24] V. Popov, I. Vovk, V. Gogoryan *et al.*, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *arXiv preprint*, vol. arXiv:2109.13821, 2022.

[25] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast griffin-lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[26] K. Kumar, R. Kumar, T. de Boissiere *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[27] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[28] N. Chen, Y. Zhang, H. Zen *et al.*, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint*, vol. arXiv:2009.00713, 2020.

[29] Y. Xie, R. Fu, Z. Wen *et al.*, "Generalized source tracing: Detecting novel audio deepfake algorithm with real emphasis and fake dispersion strategy," in *Interspeech 2024*, 2024, pp. 4833–4837.

[30] N. Klein, T. Chen, H. Tak *et al.*, "Source tracing of audio deepfake systems," in *Interspeech 2024*, 2024, pp. 1100–1104.

[31] G. Eren and The Coqui TTS Team, "Coqui TTS," Jan. 2021. [Online]. Available: https://github.com/coqui-ai/TTS

[32] K. Lee, "Comprehensive-tacotron2," https://github.com/keonlee9420/Comprehensive-Tacotron2, 2021.

[33] ——, "DiffGAN-TTS," Feb. 2022. [Online]. Available: https://github.com/keonlee9420/DiffGAN-TTS

[34] H. You, "Multispeaker GlowTTS," https://github.com/CODEJIN/Glow_TTS, Jan. 2020.

[35] ——, "SPEECHSPLIT," https://github.com/CODEJIN/SPEECHSPLIT, Jan. 2020.

[36] T.-V. Trinh, D. Yeung, I. Vovk *et al.*, "Diffusion-Based Any-to-Any Voice Conversion," Jan. 2023. [Online]. Available: https://github.com/trinhtuanvubk/Diff-VC

[37] Descript Inc., "Official repository for the paper MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," Jan. 2019. [Online]. Available: https://github.com/descriptinc/melgan-neurips/

[38] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[39] H. You, "Parallel WaveGAN," Jan. 2020. [Online]. Available: https://github.com/CODEJIN/PWGAN_Torch

[40] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[41] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," in *In Proc. ICML*, 2023.

[42] "Jiwer," https://pypi.org/project/jiwer/, 2019, apache-2.0 License.

[43] A. Conneau, A. Baevski, R. Collobert *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech 2021*, 2021, pp. 2426–2430.

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

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

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