INTRODUCING THE MNW BENCHMARK FOR AI FORENSICS

Thomas Roca
Microsoft AI for Good

Marco Postiglione
Northwestern University

Chongyang Gao
Northwestern University

Isabel Gortner
Northwestern University

Zuzanna Wojciak
WITNESS

Pengce Wang
Microsoft AI for Good

Mahsa Alimardani
WITNESS

Shirin Anlen
WITNESS

Kevin White
Microsoft AI for Good

Juan Lavista Ferres
Microsoft AI for Good

Sarit Kraus
Bar-Ilan University

Sam Gregory
WITNESS

V.S. Subrahmanian
Northwestern University

ABSTRACT

We introduce the *Microsoft-Northwestern-WITNESS* deepfake detection benchmark, a dataset to help evaluate and improve algorithms to detect AI-generated audio, video, and image content. This dataset consists of more than 50,000 artifacts (images, videos and audio files) all generated by us. It also includes real-world examples of AI-manipulated or suspicious media encountered by journalists and human rights defenders globally, annotated by experts, ensuring that the benchmark reflects the practical and high-stakes conditions under which detection tools are needed. The MNW dataset will be periodically updated to cover as many generators as possible. Our dataset also contains adversarial examples produced with state-of-the-art attacks. This project is a collaborative effort, and we encourage developers and providers of generative-AI models to help us keep this dataset always up-to-date. **This dataset is intended for evaluation purposes and cannot be used for training or for commercial purposes.** As we cannot prevent detector builders to do so, we recommend entities willing to purchase a detection solution not to use our dataset to evaluate commercial detection tools. Our objective is to set a high bar for developers and increase the reliability of detection systems.

1 Introduction

In recent years, generative AI - especially diffusion-based models - have advanced rapidly. With the help of cloud computing, the ability to create realistic AI-generated content such as images, audio, and video is now accessible to anyone through a simple text prompt, reference image, or voice input. This technology has been widely embraced by artists, content creators, and communication professionals to enhance creativity and productivity. However, it is also increasingly used for harmful purposes, including misinformation and disinformation, propaganda, fraud and online harassment. Figure 1 shows some sample AI-Generated images from 2025. While image manipulation techniques have existed long before the digital age, the speed and ease of AI-driven content generation is unlike anything seen before.

Figure 1: Examples of AI-generated images circa 2025

AI-generated media and deepfake detection has been an active research area for almost a decade. However, the past three years have seen rapid breakthroughs in generative models, along with a surge in the number of providers of generative models. What was once a niche domain with a few techniques (GANs, VAEs) has now evolved into a fast-moving landscape, with new generative models and updates released regularly. *Midjourney*, for instance, launched its seventh image generation model earlier this year. Despite their realism, generative models leave behind subtle, model-specific artifacts. While some detectors can generalize to unseen new generators, our findings show that high and consistent detection performance also depends on exposure to a diverse range of generators during training. In this context, it has become challenging to evaluate the performance of AI detectors.

Effective benchmarks should also reflect real-world conditions including compressed, low-resolution, and contextually complex media, as well as the diverse linguistic and regional contexts in which journalists and fact-checkers operate. Detection systems must be designed for technical performance as well as for the real-world, where stakes and potential harms can be high.

1.1 The Need for Comprehensive and Up-to-date benchmarks

Historically, evaluation of deepfake detection models was based on large datasets opened-up during 'detection challenges'. These datasets typically had a lot of depth but little breadth. Table4 in the appendix shows a list of deepfake benchmarks. Such benchmarks were suitable for the previous era when GANs were the dominant method for media generation. However, these benchmarks are not up to the challenge posed by the new generative AI landscape and the evolving types of risks they pose. Such risks include scams, fraud, non-consensual intimate image and video generation, and misinformation and disinformation at scale.

Though the evaluation of detection tools should reflect the evolution of the generative AI landscape, the vast majority of benchmarks available today still mostly cover GAN-based images and deepfake videos, while the few that contain media generated with diffusion-based methods only gather samples from a few open-source generative models. For

example, Pellegrini and al. 2025 [45] cover only 8 diffusion-based image generators, half of which are versions of *Stable Diffusion* and exclude the last two updates *Stable Diffusion 3* and *3.5*.

Our novel Microsoft-Northwestern-Witness (MNW) benchmark introduces a comprehensive dataset of over 50K artifacts designed to cover a broad range of AI content generators including all three types of artifacts: audio, video, and image. Our goal is to keep our dataset up-to-date by continuously integrating samples from newly released models. We believe that effective evaluation of detection systems does not require millions of examples from a few generators, but rather a relatively small number of samples from as broad a set of generators as possible. Our research shows that generative models leave behind subtle yet distinct artifacts, and that cross-generator detection remains a significant challenge for most detectors. The MNW dataset can be downloaded from: `https://github.com/nsail-lab/MNW` for non-commercial use.

We conceive this project as a collaborative endeavor and invite both the research community and model developers to support this initiative by contributing samples from new generative models, helping to keep this benchmark up-to-date across three modalities: audio, images, and video.

1.2 The Need for Benchmarks that Contain Real-World and Adversarial Examples

A few years back, concerns around Generative AI primarily centered on impersonation facilitated by deepfake videos. At the time, the threat was relatively constrained by the technical expertise and computing power required to produce such content. While real-world incidents occasionally made headlines, they remained rare. Today, generative models are accessible through user-friendly interfaces and simple natural language prompts, dramatically lowering the barrier to entry. AI-generated media are now widely disseminated on social networks, often in highly compressed, resized, or cropped form. Malicious actors increasingly employ techniques to obscure the synthetic origin of media — such as by inserting AI-generated frames into real footage, deliberately reducing resolution to mask artifacts, or injecting noise into audio. Moreover, skilled adversaries can deploy adversarial attacks to deceive detection systems and conceal the synthetic nature of images, audio, or video.

To ensure the effectiveness of AI detection tools in real-world scenarios, evaluation datasets must include both authentic and adversarial examples that reflect the conditions under which generative media are likely to be used.

In all, the initial version of the MNW benchmark includes over 50,000 artifacts comprising 40,000 deepfake audios, and over 11,000 deepfake images and videos.

2 The benchmark

With the *Microsoft-Northwestern-WITNESS* benchmark, we fill the gaps previously mentioned, bringing our 3 founding organizations' specific expertise.

Microsoft's *AI for Good Lab* is dedicated to developing and applying artificial intelligence to address real-world challenges that advance global well-being. Its initiatives span a wide range of impactful areas, including AI-powered, solar-enabled systems for monitoring biodiversity in remote regions, efforts to protect the Amazon rainforest, strategies to combat malnutrition, innovations in ear health, and tools for identifying at-risk communities. The lab also conducts research in the field of AI-forensics to counter the misuse of generative AI and improve the reliability of AI detection tools.

Northwestern University's Security and AI Lab is a leader at the intersection of AI and global security issues. Since July 8 2024, it has provided the Global Online Deepfake Detection System (GODDS) as a global public good for use by journalists[48]. GODDS has been used by journalists from numerous outlets around the world including Agence France Press, BBC, CNN, New York Times, Politifact, Thomson Reuters, Wall Street Journal, and many others. Likewise, it has provided the Northwestern Terror Early Warning System (NTEWS) as a global public good for governments and companies, forecasting terror attacks by 6 major groups. Users of NTEWS reports include governmental entities such as the UN, DHS, TSA, FEMA, Ministries of Defence of India and The Netherlands, Interpol, as well as corporations such as Ford, United Airlines, Microsoft, and others. Several other projects focus on securing the world by protecting children, detecting malware, detecting social bots and mis/disinformation, detecting phishing, scam detection, and more.

WITNESS' Deepfakes Rapid Response Force (DRRF) was the first global mechanism for forensic analysis of suspected deepfakes, providing real-time support to frontline journalists and fact-checkers while delivering cutting-edge detection training. Data from the DRRF have been used to develop the TRIED: Truly Innovative and Effective AI Detection Benchmark [2], and now contribute to this wider MNW Benchmark.

2.1 Real-world examples

As part of the MNW benchmark, WITNESS contributed to the "AI Media In-the-Wild" dataset [64], based on submissions from partners to the Deepfakes Rapid Response Force (DRRF). The DRRF connects frontline fact-checkers, journalists, and civil society actors with leading media forensics and deepfake detection experts to provide rapid evidence-based analysis of potentially AI-manipulated or generated content threatening democracy and human rights.

The dataset consists of videos, audio recordings, and images submitted to and analyzed by DRRF experts, reflecting the diversity of real-world cases encountered in high-stakes contexts. It is structured using three practical labels:

- `likely-manipulated`: Expert analysis found evidence of AI manipulation indicating the content was likely AI-generated or altered.
- `likely-authentic`: Expert analysis did not find evidence of AI manipulation, indicating the content was likely authentic.
- `inconsistent`: Expert analysis could not deliver a conclusive result due to factors such as low quality, high compression, or language/context gaps.

Examples include:

- **Likely authentic:** A022 [69], A052 [75], A064 [72]
- **Likely manipulated:** M0391 [67], M056 [65], M066 [68], M067 [70], M0842 [73]
- **Inconsistent:** I043 [74], I082 [66], I085 [71]

Figure 2: In-the-wild examples gathered through the work of DRRF

This dataset complements the WITNESS TRIED Benchmark [2], which provides a socio-technical lens for evaluating AI detection tools by prioritizing real-world applicability, contextual relevance, accessibility, and fairness. By incorporating "in-the-wild" examples, the MNW benchmark ensures that detection models are evaluated under the practical conditions faced by journalists, fact-checkers, and human rights defenders, where compressed, low-quality, and contextually complex media are the norm.

Through this contribution, WITNESS aims to support the development of reliable, explainable, and effective detection tools that strengthen global information integrity and public trust in an era of rapidly evolving generative AI threats.

2.2 AI-Generated Audio

As part of the MNW Benchmark, the Northwestern University Security and AI Lab (NSAIL) developed a total of 40,000 synthetic audio clips. As deceased individuals are not subject to Institutional Review Board protections, we selected a set DI of 206 deceased individuals di, spanning a mix of male and female subjects. We also selected 40 well-known audio generation methods from 2017-2025. These methods included many recent generative methods (e.g., 9 from 2025, 7 from 2024). We randomly sampled 250 transcripts per method from the VCTK corpus [79], a dataset widely used in voice synthesis benchmarks [77, 60]. For each method, we generated synthetic audio clips using these transcripts and a reference voice. When possible, we used speaker identities officially released by the authors of the corresponding toolkit — otherwise, we selected a voice from our DI set.

The resulting audio clips vary in duration, with a mean of 3.625 seconds, a median of 2.613 seconds, and a standard deviation of 6.859 seconds. A mean UMAP projection of Whisper embeddings [50], visualizing the distribution of audio generators, is shown in Figure 5. Table 6 in the Appendix shows the specific generators used, along with appropriate citations.

This led to a total of $40 \times 250 = 10,000$ audio deepfake samples. We then applied three types of perturbations to each audio sample: (1) additive Gaussian noise, (2) background noise randomly sampled from one of the 2,000 environmental recordings in the ESC-50 dataset [47], and (3) time-stretching of the audio signal [76]. This results in the creation of another 30,000 audio deepfake samples, leading to a total of 40,000 audio deepfakes in all in the current version of the MNW dataset.

The audio component of the MNW dataset introduces several key innovations compared to prior benchmarks (e.g., ASVspoof2019 [60], WaveFake [16], In The Wild [40], CodecFake [34]; see Table 5 in the Appendix for a more comprehensive list). First, it offers unparalleled breadth, encompassing synthetic audio generated using 40 distinct voice generation methods including very recent generative models. Second, it captures a wide range of speaker variability by leveraging reference voices from our set of 206 deceased individuals (DI), as well as publicly released

reference voices from the toolkits used. Third, the dataset uniquely accounts for adversarial perturbations that could be introduced by malicious actors to evade detection, such as added background noise, Gaussian noise, and time-stretching transformations.

2.3 AI-generated Images and video

Rather than including outdated models for which samples are readily available online, the MNW Dataset focuses on recent diffusion-based generators for AI-images, AI-video, and SOTA deepfake video architectures (GAN or diffusion-based). Although we make a distinction between AI-video and deepfakes, these two categories have started to merge with the progress achieved by AI-avatar systems such as Heygen, Vasa1, and others. In the MNW dataset, deepfakes are a specific type of AI-video specifically designed to impersonate someone. It involves a narrator and lip synchronization to an audio file. Most deepfake systems relies on GAN generators to reproduce lips, jaws, etc. although diffusion-based generators have also been also introduced (e.g. diff2lip).

At the time of the launch of this dataset (July 2025), we are making available samples from:

- 43 diffusion-based generators (250 images per generator) - see Table1
- 13 diffusion-based video generators (10 videos per generator) - see Table2
- 14 deepfake generators, diffusion and GAN-based, without audio track (10 per generator)- see Table3

We chose to limit the number of examples per generator for several reasons:

- Overall we provide 10,750 images (when adding adversarial examples) and 270 videos (i.e., more than 30,000 frames) which we consider enough to get a good idea of the performance a detectors.
- We want to avoid the use of the MNW dataset as a training set and wish to limit the dataset size to reduce this risk.
- We want to keep this dataset up-to-date. Limiting the number of examples makes it realistic to gather media from generators only accessible behind a user interface (no API or model weights accessible).

To illustrate the wide variety of the AI-image generators we provide samples from, we mapped the mean Vision Transformers CLS embedding for each generator (see Figure4 in Appendix).

2.3.1 The specific case of image inpainting

We chose to list *Flux inpainting* and *Stable Diffusion XL inpainting* as specific generators, although they are not really so. The outcome leaves distinct artifacts, as only the pixels from the inpainting mask is AI generated. Their signature on an image will then be different compared to images that are entirely AI-generated. Some inpainting processes can also include post-processing. These make inpainting a specific category that are generally harder to detect. These specific types of AI generated images can be used for malicious purposes such as inserting persons, objects, etc. into real scenes.

Figure 3: Inpainting generation pipeline

2.3.2 Adversarial examples of AI generated images

Detection algorithms can be fooled using adversarial techniques. Although these techniques often target a specific detector by approximating its decision boundary, our research shows that such attacks are partially transferable. By including these examples in the MNW evaluation set, our goal is to encourage model developers to red-team their

AI image generators - as of July 2025		
Adobe Firefly v2	HyperSD bytedance Flux	OpenAI GPTo4 image
Adobe Firefly v3	HyperSD bytedance SD3	Pixart alpha XL
Adobe Firefly v4	Ideogram	Playground-v2.5Ae
Amazon Titan v2	Kandinsky	PlaygroundAI
Aura flow	Krea 1	Recraft v3
Baidu	Kuaishou Kolors	Reve AI
ByteDance Seedream v3	Luma Photon	Stable diffusion XL img reference
Civitai v6	Lumina	Stable diffusion XL inpainting
Flux Dev	Meta Imagine	Stable diffusion v2.1
Flux Kontext Pro	Midjourney v5	Stable diffusion v3
Flux inpainting	Midjourney v6	Stable diffusion v35 Large
Flux pro1.1	Midjourney v7	Stable diffusion v35 large turbo
Google Imagen3	Nvidia Sana	Ultrapixel
Google Imagen4	OpenAI Dalle2	Wuerstchen
HunyuanDiT	OpenAI Dalle3	

Table 1: AI Image Generators (diffusion-based) used to produced image samples available in the MNW dataset

detection systems. So far, we have added 250 examples of diverse, state-of-the-art white-box attacks that have demonstrated transferability. We also plan to expand this section of the dataset over time and welcome contributions in this area.

2.3.3 AI videos and Deepfake videos

In this dataset, we distinguish between *AI videos* and *Deepfake videos*. We define an AI video as a video that only contains frames that are entirely AI-generated. These videos do not necessarily aim to mislead viewers. Examples of such generators are OpenAI Sora, Google Veo, Kling, etc. - see in Table 2 the list of AI video generators we provide samples from. We chose to provide only 10 videos per generator (which translates into approximately 1,200 frames per generator) which is enough to get a sense of the performance of a detection system, but not enough to train one only using those, while allowing to produce examples from as many generators as possible.

AI video generators - as of July 2025	
Adobe Firefly video	Midjourney Video
CogVideoX-2B	OpenAI Sora
Google VEO3	Pika 1.5
Kling 1.6	Pika 2.1
Kling 2.0 Master	Pika 2.2
Kling 2.1	Runway Gen4
Kling 2.1 Master	

Table 2: AI Video Generators (diffusion-based) used to produced video samples available in the MNW dataset

Deepfake videos, on the other hand, are designed to impersonate a narrator. They typically use voice cloning for the audio track and AI models to synchronize lips, jaw, or facial movements. GANs and diffusion models are often employed to generate facial pixels, which are then substituted into the original real video - see in Table 3 the list of Deepfake systems this dataset covers. The artifacts left by such manipulations differ significantly from those resulting from the generation of an entire frame or image.

With recent advances in AI-avatars, the boundary between AI videos and Deepfakes is becoming increasingly blurred. The addition of sound and voice capabilities in the latest release of Google's Veo further contributes to this convergence.

We plan to expand this section of the dataset to include AI-avatar videos and welcome contributions in this area.

NB. As AI-generated audio is covered in another section of the dataset, we have removed the audio tracks from both the AI and Deepfake video samples.

Deepfake video generators - as of July 2025	
Diff2lip	SadTalker video v2
Echo mimic	Vasa 1
Heygen v1 (lip sync services)	Video retalking
MuseTalk	Wav2lip
RaskAI (lip sync services)	Wav2lip GFPGAN
SadTalker video	Wav2lipHQ ESRGAN

Table 3: Deepfake video Generators (GAN and diffusion-based) used to produced deepfake video samples available in the MNW dataset

3 Conclusion

Our initiative has brought together a diverse dataset spanning audio, video, and image modalities, including adversarial and real-world examples to contribute to the MNW Benchmark.

By incorporating the WITNESS "AI Media In-the-Wild" dataset, the benchmark ensures that evaluations reflect the practical conditions faced by journalists, fact-checkers, and human rights defenders, aligning with the TRIED Benchmark's call for a socio-technical lens in detection evaluation.

By incorporating the Northwestern Audio Deepfake Dataset consisting of 40K audios, the benchmark ensures the availability of deepfake audios generated by a wide variety of *recent* audio deepfake generators, as well as some common methods used by adversaries to evade detection.

Likewise, the Microsoft visual dataset ensures the availability of AI and deepfake videos and images generated by a wide variety of *recent* generators.

Please note that the MNW dataset is intended for evaluation purposes and cannot be used for training or for commercial purposes.

4 Appendix

Table 4: Benchmark dataset for Deepfake videos

Datasets	Release Date	Real Videos	Fake Videos	Fake Video Generation Technique
UADFV [29]	Nov. 2018	48	48	FakeApp Application
DeepfakeTIMIT [24]	Dec. 2018	0	620	faceswap-GAN
FaceForensics++ [53]	Jan. 2019	1000	3000	face2Face,FaceSwap,CG-manip.
Deepfake Detection Challenge [13]	Oct. 2019	1131	4119	faceswap: DFAE, MM/NN,NTH,..
Celeb-DF [30]	Nov. 2019	408	795	deepfake
DFD [15]	2019	360	3000	deepfakes
DeeperForensics-1 [20]	June 2020	10000	50000	DF-VAE
WildDeepFake [83]	Oct. 2020	3805	3509	Unknown, collected from the web
Celeb-DF (v2) [31]	2020	590	5639	deepfake
DeepFake Game Competition - DFGC [46]	2021	-	-	deepfake - CelebDF-v1-v2
OpenForensics [27]	2021	45,473	70,325	GAN based
KoDF [25]	2021	62,166	175,776	DF-VAE
FakeAVCeleb [21]	2021	500	19,500	faceswap: FSGAN,sv2tts,wav2lip
HiFiFace [61]	2021	-	1000	HifiFace

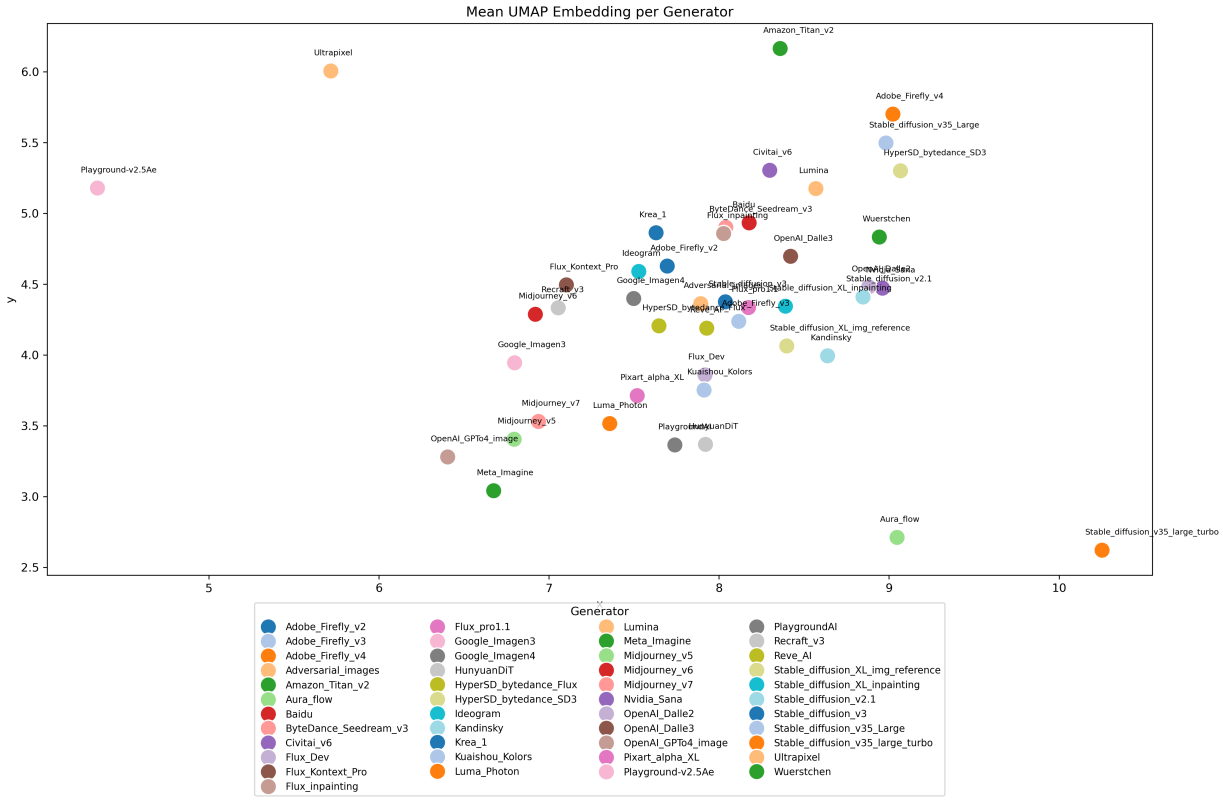

Figure 4: Cluster analysis: mean UMAP ViT CLS embeddings by Image generator

Table 5: Benchmark datasets for Deepfake audios

Datasets	Release Date	Real Audios	Fake Audios	Deepfake Type
ASVspoof2015 [78]	Sep. 2015	16,651	246,500	TTS, VC
FoR [51]	Oct. 2019	108,256	87,285	TTS
ASVspoof2019-LA [60]	Nov. 2019	10,256	90,192	TTS, VC
ASVspoof2021-LA [33]	Sep. 2021	14,816	133,360	TTS, VC
ASVspoof2021-DF [33]	Sep. 2021	14,869	519,059	TTS, VC
FMFCC-A [82]	Oct. 2021	10,000	40,000	TTS, VC
WaveFake [16]	Nov. 2021	0	117, 985	TTS
InTheWild [40]	Sep. 2022	19,963	11,816	Unknown
TIMIT-TTS [54]	Apr. 2023	0	5,160	TTS
MLAAD [41]	Jan. 2024	0	76,000	TTS
CodecFake [34]	Jun. 2024	44,242	44,242	codec
CFAD [36]	Oct. 2024	38,600	77,200	TTS, VC

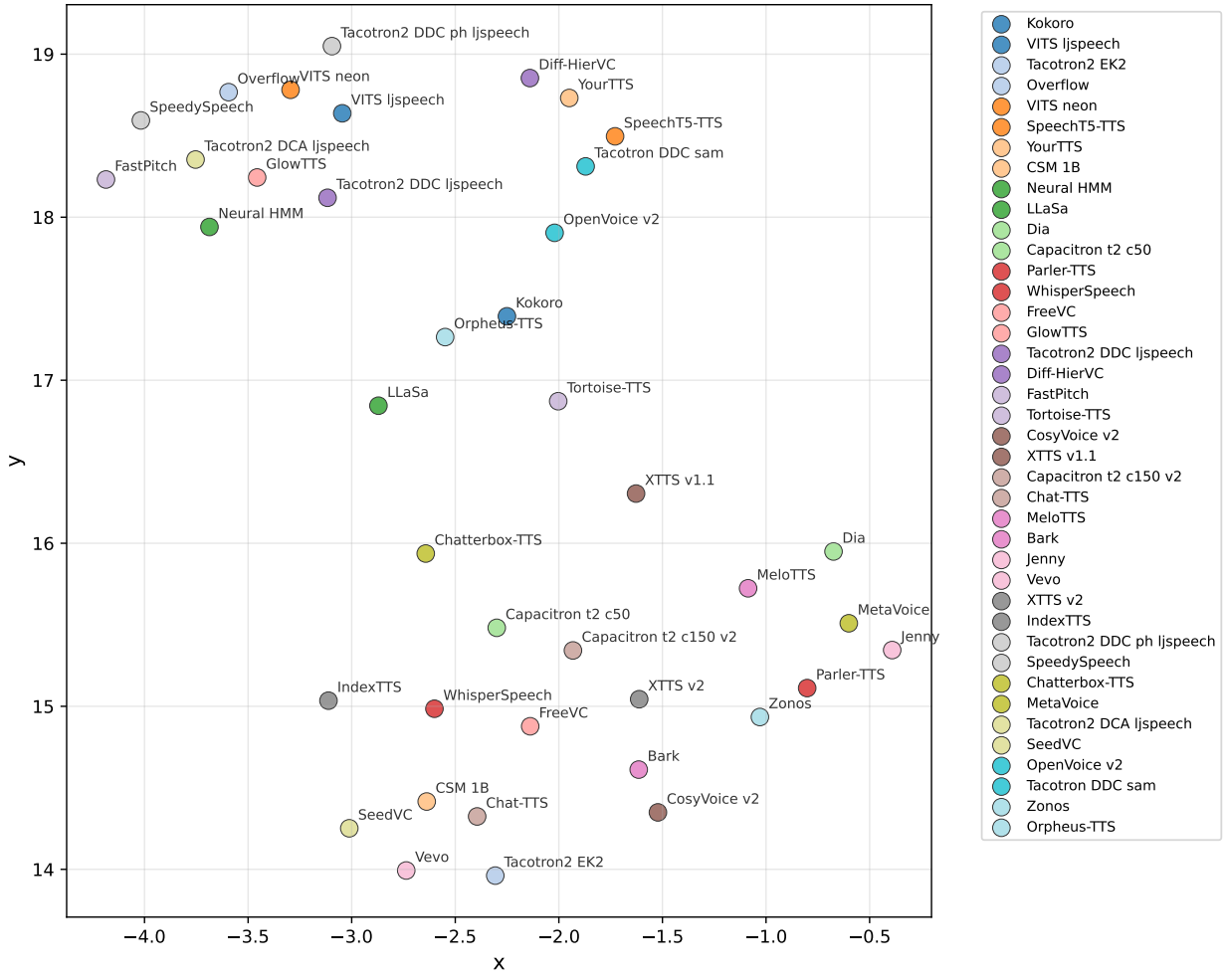

Figure 5: Cluster analysis: Mean UMAP Whisper Embeddings by Audio Generator

References

[1] 2noise. Chattts: A generative speech model for daily dialogue. `https://github.com/2noise/ChatTTS`, 2024. GitHub repository; accessed July 1, 2025.

Table 6: **Deepfake audio generators** - as of July 2025

Method	Release Year
Chatterbox-TTS [52]	2025
Dia [43]	2025
Zonos [84]	2025
Vevo [81]	2025
Orpheus-TTS [7]	2025
LLaSa [80]	2025
Kokoro [18]	2025
IndexTTS [12]	2025
CSM-1B [56]	2025
Chat-TTS [1]	2024
Parler-TTS [35]	2024
Melo-TTS [42]	2024
XTTS (v2) [8]	2024
MetaVoice [39]	2024
SeedVC [32]	2024
CosyVoice-v2 [14]	2024
Tortoise-TTS [6]	2023
Overflow [37]	2023
OpenVoice-v2 [49]	2023
XTTS (v1.1) [11]	2023
WhisperSpeech [63]	2023
Diff-HierVC [10]	2023
FreeVC [28]	2023
Bark [58]	2023
Neural HMM [38]	2022
SpeechT5-TTS [3]	2022
Jenny [11]	2022
VITS (uses Neon [44]) [23]	2021
VITS [23]	2021
FastPitch [26]	2021
GlowTTS [22]	2020
SpeedySpeech [59]	2020
YourTTS [9]	2020
Capacitron (embedding capacity: 50) [4]	2020
Capacitron (embedding capacity: 150) [4]	2020
Tacotron 2 [57] (training set: LJSpeech [19]; uses Dynamic Convolutional Attention [5])	2020
Tacotron 2 [57] (training set: EK1)	2018
Tacotron 2 [57] (training set: LJSpeech [19]; uses Double Decoder Consistency (DDC) [17]	2018
Tacotron 2 [57] (training set: LJSpeech [19]; uses DDC [17]; uses phonemes)	2018
Tacotron [62] (training dataset: SAM [55])	2017

[2] Shirin Anlen and Zuzanna Wojciak. Tried: Truly innovative and effective ai detection benchmark, developed by witness, 2025.

[3] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, May 2022.

[4] Eric Battenberg, Soroosh Mariooryad, Daisy Stanton, R. J. Skerry-Ryan, Matt Shannon, David Kao, and Tom Bagby. Effective use of variational embedding capacity in expressive end-to-end speech synthesis. *CoRR*, abs/1906.03402, 2019.

[5] Eric Battenberg, R. J. Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom

Bagby. Location-relative attention mechanisms for robust long-form speech synthesis. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6194–6198. IEEE, 2020.

[6] James Betker. Better speech synthesis through scaling. *CoRR*, abs/2305.07243, 2023.

[7] canopyai. Orpheus-tts: Towards human-sounding speech. `https://github.com/canopyai/Orpheus-TTS`, 2025. GitHub repository; accessed July 1, 2025.

[8] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: a massively multilingual zero-shot text-to-speech model. In Itshak Lapidot and Sharon Gannot, editors, *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA, 2024.

[9] Edresson Casanova, Julian Weber, Christopher Dane Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir A. Ponti. Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR, 2022.

[10] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. In Naomi Harte, Julie Carson-Berndsen, and Gareth Jones, editors, *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 2283–2287. ISCA, 2023.

[11] coqui-ai. Coqui tts. `https://github.com/coqui-ai/TTS`, December 2022. GitHub repository; accessed June 30, 2025.

[12] Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*, 2025.

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[14] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117, 2024.

[15] Nick Dufour and Andrew Gully. Google ai: Contributing data to deepfake detection research, September 2019.

[16] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[17] Eren Gölge. Solving attention problems of tts models with double decoder consistency. Coqui.ai Blog, June 2020. Accessed: 2025-07-01.

[18] hexgrad. Kokoro. `https://github.com/hexgrad/kokoro`, 2025. GitHub repository; accessed July 1, 2025.

[19] Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[20] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898, 2020.

[21] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[22] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[23] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR, 2021.

[24] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[25] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021.

[26] Adrian Lancucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6588–6592. IEEE, 2021.

[27] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10117–10127, 2021.

[28] Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023.

[29] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018.

[30] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df (v2): a new dataset for deepfake forensics [j]. *arXiv preprint arXiv*, 2019.

[31] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[32] Songting Liu. Zero-shot voice conversion with diffusion transformers. *CoRR*, abs/2411.09943, 2024.

[33] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2507–2522, 2023.

[34] Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. Codecfake: An initial dataset for detecting llm-based deepfake audio. In Itshak Lapidot and Sharon Gannot, editors, *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA, 2024.

[35] Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

[36] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, and Ruibo Fu. CFAD: A chinese dataset for fake audio detection. *Speech Commun.*, 164:103122, 2024.

[37] Shivam Mehta, Ambika Kirkland, Harm Lameris, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Overflow: Putting flows on top of neural transducers for better TTS. In Naomi Harte, Julie Carson-Berndsen, and Gareth Jones, editors, *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4279–4283. ISCA, 2023.

[38] Shivam Mehta, Éva Székely, Jonas Beskow, and Gustav Eje Henter. Neural HMMS are all you need (for high-quality attention-free TTS). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7457–7461. IEEE, 2022.

[39] metavoiceio. metavoice-src: Foundational model for human-like, expressive tts. `https://github.com/metavoiceio/metavoice-src`, 2024. GitHub repository; accessed July 1, 2025.

[40] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In Hanseok Ko and John H. L. Hansen, editors, *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2783–2787. ISCA, 2022.

[41] Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. MLAAD: the multi-language audio anti-spoofing dataset. In *International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, June 30 - July 5, 2024*, pages 1–7. IEEE, 2024.

[42] mysell-ai. Melotts: High-quality multi-lingual text-to-speech library by myshell.ai. support english, spanish, french, chinese, japanese and korean. `https://github.com/myshell-ai/MeloTTS`, 2024. GitHub repository; accessed July 1, 2025.

[43] nari-labs. Dia: A tts model capable of generating ultra-realistic dialogue in one pass. https://github.com/nari-labs/dia, 2025. GitHub repository; accessed July 1, 2025.

[44] NeonGeckoCom. neon-tts-plugin-coqui. https://github.com/NeonGeckoCom/neon-tts-plugin-coqui, March 2025. GitHub repository; accessed July 1, 2025.

[45] Lorenzo Pellegrini, Davide Cozzolino, Serafino Pandolfini, Davide Maltoni, Matteo Ferrara, Luisa Verdoliva, Marco Prati, and Marco Ramilli. AI-GenBench: A New Ongoing Benchmark for AI-Generated Image Detection, April 2025. arXiv:2504.20865.

[46] Bo Peng, Hongxing Fan, Wei Wang, Jing Dong, Yuezun Li, Siwei Lyu, Qi Li, Zhenan Sun, Han Chen, Baoying Chen, Yanjie Hu, Shenghai Luo, Junrui Huang, Yutong Yao, Boyuan Liu, Hefei Ling, Guosheng Zhang, Zhiliang Xu, Changtao Miao, Changlei Lu, Shan He, Xiaoyan Wu, and Wanyi Zhuang. DFGC 2021: A DeepFake Game Competition, June 2021. arXiv:2106.01217.

[47] Karol J. Piczak. ESC: dataset for environmental sound classification. In Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan, editors, *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 1015–1018. ACM, 2015.

[48] Marco Postiglione, Julian Baldwin, Natalia Denisenko, Luke Fosdick, Chongyang Gao, Isabel Gortner, Chiara Pulice, Sarit Kraus, and V. S. Subrahmanian. GODDS: the global online deepfake detection system. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 29685–29687. AAAI Press, 2025.

[49] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023.

[50] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[51] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In Corneliu Burileanu and Horia-Nicolai Teodorescu, editors, *2019 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019, Timisoara, Romania, October 10-12, 2019*, pages 1–10. IEEE, 2019.

[52] resemble-ai. chatterbox: Sota open-source tts. https://github.com/resemble-ai/chatterbox, 2025. GitHub repository; accessed July 1, 2025.

[53] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[54] Davide Salvi, Brian C. Hosler, Paolo Bestagini, Matthew C. Stamm, and Stefano Tubaro. TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection. *IEEE Access*, 11:50851–50866, 2023.

[55] Sam-Accenture-Non-Binary-Voice. non-binary-voice-files. https://github.com/Sam-Accenture-Non-Binary-Voice/non-binary-voice-files, June 2021. GitHub repository; accessed July 1, 2025.

[56] SesameAILabs. csm: A conversational speech generation model. https://github.com/SesameAILabs/csm, 2025. GitHub repository; accessed July 1, 2025.

[57] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783. IEEE, 2018.

[58] suno-ai. Bark: Text-prompted generative audio model. https://github.com/suno-ai/bark, April 2023. GitHub repository; accessed July 1, 2025.

[59] Jan Vainer and Ondrej Dusek. Speedyspeech: Efficient neural speech synthesis. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 3575–3579. ISCA, 2020.

[60] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas W. D. Evans, Md. Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.*, 64:101114, 2020.

[61] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021.

[62] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In Francisco Lacerda, editor, *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*, pages 4006–4010. ISCA, 2017.

[63] WhisperSpeech. Whisperspeech: An open source text-to-speech system built by inverting whisper. https://github.com/WhisperSpeech/WhisperSpeech, 2023. GitHub repository; accessed July 1, 2025.

[64] WITNESS. Ai media in-the-wild dataset (mnw benchmark contribution), 2025. Accessed: July 2025.

[65] WITNESS. Archived facebook video. https://web.archive.org/web/20250616100542/https://www.facebook.com/100002167747651/videos/457953553957488/?vh=e&extid=MSG-UNK-UNK-UNK-COM_GK0T-GK1C, 2025. Archived July 2025.

[66] WITNESS. Archived facebook video. https://web.archive.org/web/20250616134222/https://www.facebook.com/irak.li.zakareishvili.satadarigo.pb/videos/738168888779154, 2025. Archived July 2025.

[67] WITNESS. Archived tiktok video. https://web.archive.org/web/20250613161804/https://www.tiktok.com/@ulyublena_storinka/video/7347380104663977222, 2025. Archived July 2025.

[68] WITNESS. Archived tiktok video. https://web.archive.org/web/20250616093610/https://www.tiktok.com/@mustafali2001/video/7438149251449752850, 2025. Archived July 2025.

[69] WITNESS. Archived twitter video: Rsf sudan. https://web.archive.org/web/20230728164151/https://twitter.com/RSFSudan/status/1684941175394914304, 2025. Archived July 2025.

[70] WITNESS. Archived x/twitter post. https://archive.ph/Y72jE, 2025. Archived July 2025.

[71] WITNESS. Archived x/twitter post. https://archive.ph/ABfWs, 2025. Archived July 2025.

[72] WITNESS. Archived youtube speech clip. https://web.archive.org/web/20250613161058/https://www.youtube.com/watch?v=K1KYhenW3oM&t=596s, 2025. Archived July 2025.

[73] WITNESS. Archived youtube video. https://web.archive.org/web/20250616101755/https://www.youtube.com/watch?v=jXw_PoltZ70, 2025. Archived July 2025.

[74] WITNESS. Archived youtube video. https://web.archive.org/web/20250616135041/https://www.youtube.com/watch?v=T8RE-8OfFp4&t=624s, 2025. Archived July 2025.

[75] WITNESS. Archived youtube video: Sudan rsf. https://web.archive.org/web/20250613160222/https://www.youtube.com/watch?v=p6VnexYxwrU, 2025. Archived July 2025.

[76] Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *arXiv preprint arXiv:2404.15854*, 2024.

[77] Zhizheng Wu, Ali Khodabakhsh, Cenk Demiroglu, Junichi Yamagishi, Daisuke Saito, Tomoki Toda, and Simon King. SAS: A speaker verification spoofing database containing diverse attacks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 4440–4444. IEEE, 2015.

[78] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Md. Sahidullah, Aleksandr Sizov, Nicholas W. D. Evans, and Massimiliano Todisco. Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Top. Signal Process.*, 11(4):588–604, 2017.

[79] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). [sound], 2019.

[80] Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *CoRR*, abs/2502.04128, 2025.

[81] Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[82] Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. FMFCC-A: A challenging mandarin dataset for synthetic speech detection. In Xianfeng Zhao, Alessandro Piva, and Pedro Comesaña Alfaro, editors, *Digital Forensics and Watermarking - 20th International Workshop, IWDW 2021, Beijing, China, November 20-22, 2021, Revised Selected Papers*, volume 13180 of *Lecture Notes in Computer Science*, pages 117–131. Springer, 2021.

[83] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020.

[84] Zyphra. Zonos. `https://github.com/Zyphra/Zonos`, 2025. GitHub repository; accessed July 1, 2025.