Collecting, Curating, and Annotating Good Quality Speech deepfake dataset for Famous Figures: Process and Challenges

Hashim Ali, Surya Subramani, Raksha Varahamurthy, Nithin Adupa, Lekha Bollinani*, Hafiz Malik*

Department of Electrical and Computer Engineering, University of Michigan, USA

alhashim@umich.edu, suryasss@umich.edu, rakshav@umich.edu, adupa@umich.edu, lekhab@umich.edu, hafiz@umich.edu

Abstract

Recent advances in speech synthesis have introduced unprecedented challenges in maintaining voice authenticity, particularly concerning public figures who are frequent targets of impersonation attacks. This paper presents a comprehensive methodology for collecting, curating, and generating synthetic speech data for political figures and a detailed analysis of challenges encountered. We introduce a systematic approach incorporating an automated pipeline for collecting high-quality bonafide speech samples, featuring transcription-based segmentation that significantly improves synthetic speech quality. We experimented with various synthesis approaches; from single-speaker to zero-shot synthesis, and documented the evolution of our methodology. The resulting dataset comprises bonafide and synthetic speech samples from ten public figures, demonstrating superior quality with a NISQA-TTS naturalness score of 3.69 and the highest human misclassification rate of 61.9%.

Index Terms: Text-to-Speech, Database, political figures

1. Introduction

The last few years have seen an exceptional increase in the realism of synthesized speech [1, 2, 3, 4, 5]. This high quality of synthesized speech and the ability to distribute it through social media platforms are giving rise to manipulated information in the digital ecosystem. According to a Global Risk Report by the World Economic Forum, misinformation and disinformation are the most serious threats predicted over the next two years [6]. The report states that approximately three billion people are expected to participate in electoral polls across multiple countries over the next two years, however, the widespread use of misinformation and disinformation and the tools to disseminate may undermine the legitimacy of newly elected governments. This can result in political unrest ranging from violent protests and hate crimes to civil confrontation and terrorism.

These concerns have already manifested themselves in several high-profile incidents. In 2022, a synthetic video portrayed President Zelenskii allegedly asking for military surrender [7]. Subsequently, in 2024, a fake audio purported to be from President Biden was used in an attempt to influence voter participation in the primary elections in New Hampshire [8]. The scope of this threat became multi-national when London Mayor Sadiq Khan was targeted through fabricated audio content regarding Armistice Day observations [9]. As speech synthesis technologies advance in capability and accessibility, influential public figures face increasing exposure to voice spoofing attacks that can systematically manipulate public opinion. The development of robust audio spoofing detection systems has therefore become crucial. However, such systems require comprehensive, high-quality datasets containing authentic and synthetic speech samples from public figures. Creating these datasets presents unique challenges, particularly when dealing with prominent individuals whose voices are frequently targeted for manipulation. This underscores the urgent need for systematic approaches to building and maintaining audio spoofing detection datasets that can effectively protect high-profile individuals from voice-based impersonation attacks.

In this paper, we present a comprehensive methodology for collecting and generating synthetic speech data for high-profile political figures while documenting the challenges encountered and the solutions developed throughout the process. Our methodology emphasizes three key aspects: (1) comprehensive coverage of authentic speech in diverse real-world contexts, including political speeches, media interviews, and public statements; (2) systematic curation of high-quality audio samples that capture the distinctive vocal characteristics and speech patterns of each individual; and (3) creation of corresponding synthetic speech using multiple text-to-speech (TTS) systems to represent realistic spoofing scenarios. The process involves a carefully designed pipeline to collect and process speech samples. First, we identify and collect high-quality source material from publicly available videos, ensuring diverse speaking contexts and acoustic conditions. This is followed by rigorous preprocessing steps, including speaker diarization to isolate the target speaker, automated transcription of speech segments, audio quality assessment, and segmentation into chunks while preserving the natural flow of speech. For each authentic speech segment, we generate the corresponding synthetic speech using various TTS approaches. The resulting dataset is available on request at our lab datasets website[1].

The remainder of this paper is organized as follows. Section 2 provides a description of the existing relevant audio anti-spoofing datasets and their limitations. Section 3 describes the design considerations for audio data collection, the data collection pipeline, and the challenges faced. Section 4 describes the process for generating synthetic speech samples and the corresponding challenges. Finally, Section 5 provides the statistics of the different datasets and their quality comparisons.

2. Existing Audio Anti-Spoofing Datasets

The research community has developed various datasets to advance the field of Audio Spoof Detection. We can broadly classify these datasets into two categories, based on their speaker characteristics and intended applications. The first category, General Purpose Speaker Datasets, comprises audio data of anonymous speakers in controlled environments, focusing on

*These authors contributed equally.

[1] https://datasets.issflab.net

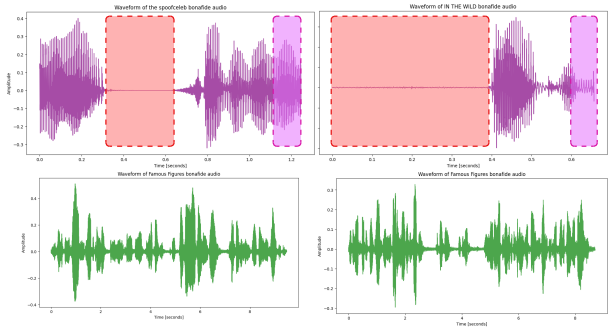

Figure 1: *Speech clips from Spoofceleb* ***(Top Left)*** *and In the Wild* ***(Top Right)*** *with extended silence* ***(Red tint)*** *, less duration and abrupt cut at the end* ***(purple tint)*** *. The bottom two plots are from Famous figures dataset with an average duration of 8 seconds and Transcription based segments*

developing generic audio spoof detection systems. The second category, Identity-Specific Datasets, addresses the challenges of protecting public figures from targeted voice spoofing attacks.

2.1. General-Purpose Speaker Datasets

General-purpose speaker datasets have played a vital role in the advancement of audio spoofing detection research by providing standardized benchmarks to evaluate detection systems. The ASVspoof Challenge series [10, 11, 12, 13] has emerged as the primary benchmark in this domain, systematically evolving to address increasingly sophisticated spoofing attacks. The latest iteration, ASVspoof 5 [13], represents a significant advancement by incorporating crowd-sourced speech data and a diverse range of deepfake attacks. It includes real-world speech extracted from the Multilingual Librispeech (MLS) English partition [14], which consists of audiobook recordings. Based on the ASVspoof series, the ASVSpoof Laundered Database (ASVSpoofLD) [15] is developed by passing audio files from the ASVSpoof19 LA eval partition through a series of laundering attacks (additive noise, reverberation, recompression, resampling, etc.), introducing additional complexity for detection systems. The DeepFake Audio Detection Dataset (DFADD) [16] and CODECFake dataset [17] represent contemporary datasets specifically designed to evaluate detection systems against recent neural TTS architectures and codec-based neural speech synthesis methods. Both datasets are derived from the VCTK corpus, which comprises high-quality speech recordings collected in a controlled laboratory environment. The Multi-Language Audio Anti-spoofing Dataset (MLAAD) [18] extends the scope to cross-lingual scenarios, addressing the increasingly global nature of audio spoofing threats.

These datasets share several key characteristics that define their utility in audio anti-spoofing research, such as controlled recording environments, standardized evaluation protocols, and balanced attack representations. However, their focus on controlled conditions, anonymous speakers, and the use of read speech limits their applicability in scenarios requiring protection of specific individuals under real-world conditions.

2.2. Identity-Specific Datasets

In contrast to general-purpose datasets, identity-specific datasets focus on protecting known individuals from targeted voice spoofing attacks. The In-The-Wild (ITW) dataset [19] represents a unique collection of real-world speech data that bridges the gap between controlled laboratory evaluations and practical applications. Unlike the previous datasets, the ITW comprises 38 hours of speech data collected from various online platforms and social media sources. A significant advancement in identity-specific datasets is represented by SpoofCeleb [20], which addresses several limitations of previous datasets by utilizing real-world data from VoxCeleb1 [21]. The authors proposed a fully automated pipeline to process VoxCeleb1 speech samples and generate the corresponding synthetic speech.

Despite the use of genuine real-world audio samples, SpoofCeleb exhibits important limitations in the context of targeted speaker protection. Most notably, the dataset's training and evaluation partitions do not share common speakers, making it more suitable for generic deepfake detection rather than protecting specific individuals from targeted attacks. Moreover, Figure 1 illustrates the data quality challenges in speech collection and their impact on the quality of the synthesis. The top waveforms are from SpoofCeleb (left) and In the Wild (right) datasets, which contain extended silences (red shading) and abrupt cuts (purple shading), which can lead to poor prosody and unnatural timing in synthetic speech. In contrast, the bottom waveforms show our Famous Figures dataset segments, which maintain an average duration of 8 seconds.

3. Dataset Design and Methodology

Our dataset design process is guided by three primary objectives: (1) ensuring high-quality authentic speech samples across various speaking contexts, (2) maintaining speaker diversity while capturing sufficient data per individual to represent their unique vocal characteristics, and (3) establishing a reproducible pipeline for data collection that can be extended to include additional public figures in the future.

3.1. Design Consideration

First, we established a criteria for selecting public figures based on three key factors: (a) frequency of public appearances, ensuring sufficient source material for data collection, (b) diversity of speaking contexts, including formal speeches, media interviews, and public statements, and (c) likelihood of being targeted for voice spoofing attacks based on their public influence. We selected 10 high-profile public figures who met these criteria. These figures include Anthony Blinken, Barack Obama, Donald Trump, JD Vance, Joe Biden, Kamala Harris, Mathew Miller, Tim Walz, Vivek Ramaswamy, and Elon Musk. Second, we followed a systematic approach to the selection of the source material. We only collected YouTube videos which have (a) minimum resolution of 720p to ensure adequate audio quality, (b) minimum video duration of 5 minutes to ensure adequate speaking patterns, (c) publication date range of 2018-2024 to ensure current speaking styles, and (d) clear speech with minimal background noise.

3.2. Data Collection Pipeline

The data collection and processing pipeline is implemented as a systematic automated workflow to ensure consistency and reproducibility. The initial data acquisition begins with a carefully curated CSV file containing YouTube links to various speeches, interviews, and public appearances of selected public figures. Each entry in the CSV file includes metadata such as the speaker's identity, start time (the time at which the target speaker starts speaking), content type (speech, interview, etc.), publication date and the YouTube URL. As described in Figure 2, the data collection pipeline consists of the following stages:

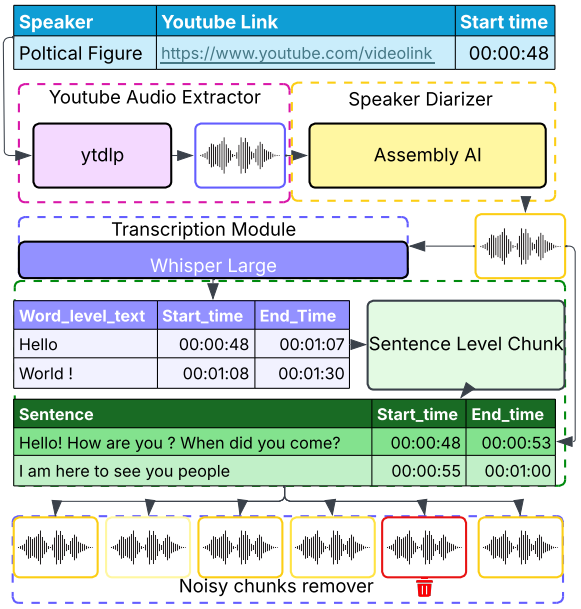

Figure 2: *Schematic diagram for real audio data collection with transcription based segmentation.*

1. **Audio Extraction:** We utilize yt-dlp[2] to directly download audio from YouTube links. It uses ffmpeg to download the best audio available in the WAV format, and resample it 16 kHz. Along with the YouTube video link, we specify the target speaker's speech starting time to trim the audio to begin from the specified timestamp, ensuring that the first speaker is the target speaker.
2. **Speaker Diarization:** We employ Assembly AI[3] to isolate segments that contain only the target speaker's voice. This step eliminates cross-talk and background speakers.
3. **Transcription Generation:** We integrate OpenAI Whisper Large Turbo[4] for transcription, which gave word-level transcripts with timestamps. We also experimented with the Google speech recognition api package and other commercial tools; however, they generated text with less accuracy and without proper punctuation.
4. **Audio Segmentation:** We utilize transcription-based segmentation that predicts word-level transcripts along with their timestamps in the utterance, as illustrated in Figure 2. This step processes word-level transcriptions with timestamps and groups them into sentence segments based on the utterance duration U, user-defined duration D, and the threshold duration T. Words are sequentially appended to a segment, and when the duration of the segment reaches $D-T$ seconds, the process searches for a punctuation mark within the interval $[D-T, D+1]$. If a punctuation mark is found within this range, the segment is finalized up to the punctuation, while the remaining words are carried over to the next segment. If no punctuation is detected, the segment extends up to $D+T$ seconds, with an additional 0.25 seconds of silence appended at the end. The last few words are discarded, if their cumulative duration is less than $D-2T$ seconds. In addition, incomplete sentences are discarded if the total segment count exceeds $\left(\frac{U}{D}-10\right)$.
5. **Quality Control:** Our initial segmentation strategy involved abrupt cutting at fixed intervals of n seconds or minutes. For example, a 30-minute utterance would be divided into 300 segments of 6 seconds each. Through continuous experimentation, we transitioned to silence- or speaker-pause-based segmentation, as described in SpoofCeleb [20]. In this approach, if silence extends beyond 500 ms, the utterance is segmented at that point. We evaluated the output audio segments based on specific criteria, such as silence duration, sentence boundary completion, and SNR range. With the updated segmentation approach, we achieved segments with maximized voiced portions, completed sentence boundaries, and naturally reduced noise.

[2]https://github.com/yt-dlp/yt-dlp
[3]https://www.assemblyai.com/
[4]https://github.com/openai/whisper/

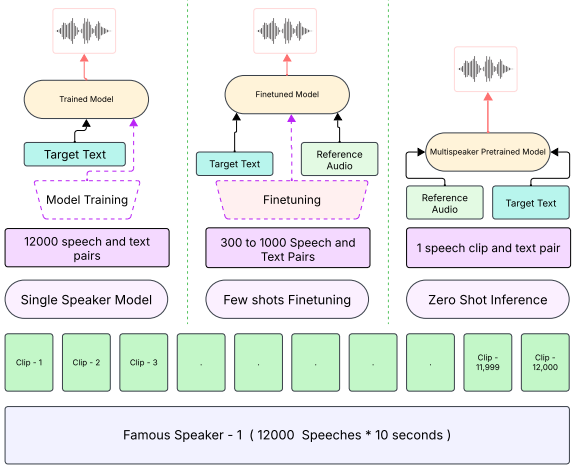

Figure 3: *Schematic diagram of speech production.*

4. Synthetic Speech Generation

The evolution of synthetic speech has progressed from supervised neural models to self-supervised learning (SSL). Although early models produced robotic-sounding speech despite extensive training data, SSL-based methods significantly improved quality through large-scale pre-training. The recent integration of Audio Language Models (ALMs) has further enhanced prosody and expressiveness.

4.1. Synthesis Pipeline

We first trained a total of 10 TTS models for each speaker using the audio samples and their corresponding transcriptions generated through the data collection pipeline in Section 3.2. After that, we used the transcripts and the trained models to generate synthetic speech. From our exploration, we have identified three primary approaches to train TTS models.

4.1.1. Speaker-Specific Training

This approach involves training a model exclusively for a single speaker, which requires at least 24 hours of speech data (approximately 11,000 to 13,000 audio samples). The training process includes creating file lists using bonafide audio samples and their corresponding transcriptions, as mentioned in Figure 3. Training is performed using high-performance GPUs, such as three NVIDIA A100 GPUs, and typically takes five days per speaker. We train StyleTTS2 [22] only using this approach. The synthetic speech generated through this approach exhibited limitations in prosody and naturalness, which led us to explore more efficient alternatives.

Table 1: *Statistical overview of audio deepfake datasets*

Dataset	#Speakers	Bonafide Utterances	Synthetic Utterances	Duration (hrs)	Speaking Contexts	Synthesis Methods
ASVspoof19 LA	107	12,483	108,978	∼100	Read speech	A01-A19 (TTS & VC)
ASVspoof 5	585	148,656	423,740	∼570	Read speech	44 TTS, VC systems
DFADD	109	44,455	163,500	∼50	Read speech	Diffusion- and Flow-matching TTS
CodecFake (EN)	110	44,242	269,903	∼312	Read speech	7 Neural Audio Codecs
SpoofCeleb	1,251	248,000	2.5M+	∼400	public appearances	23 TTS systems
In-The-Wild	54	19,963	11,816	∼38	public appearances	Unknown
Famous Figures	10	26,500	265,000	∼590	public appearances	10 Open-source TTS models

Table 2: *NISQAv2 prediction for in the wild datasets*

Dataset type	Avg. Naturalness	Fake Miss-Rates(%)
ASVspoof19 LA	2.99	25
CodecFake (EN)	3.41	57.5
DFADD	3.39	24.4
MLAAD	3.53	34.8
In the wild	2.80	52.5
Spoof celeb	3.06	-
Famous Figures	**3.69**	**61.9**

4.1.2. Few-Shot Fine-Tuning

In this approach, a model pre-trained for multiple speakers is adapted to a specific speaker using 1 to 3 hours of its data through fine-tuning. We fine-tuned XTTSv2 [23] and StyleTTS2 [22] for all speakers. This approach significantly improved speech quality by effectively transferring prosody and emotional features from the multi-speaker model.

4.1.3. Zero-Shot Synthesis

In this approach, a large-scale model pre-trained for multiple speakers is adapted to a specific speaker using only a single reference audio and text pair. Although models like XTTSv2 and StyleTTS2 have zero-shot capabilities, they struggled to match the reference speaker's voice accurately. However, recent models integrating ALM based architectures have drastically improved the performance. From this category, we have generated synthetic speech for all speakers using F5TTS[24], E2TTS[25], FishSpeech [26], SSRSpeech [27], MaskGCT [28], CozyVoice2 [29], LLASA,[30] and Zonosv0.1.

4.2. Speech Synthesis Challenges and Solutions

The development of synthetic speech for political figures presented unique challenges that required iterative solutions. We discuss our progress through multiple approaches, highlighting both the challenges encountered and the solutions implemented.

4.2.1. Challenges

We faced significant obstacles in our initial attempts to use political speech recordings to train TTS models. We faced fundamental limitations with Signal-to-Noise Ratio (SNR) measurements for publicly available recordings (12.12dB), which is substantially below the 30.25 dB benchmark established by standard TTS datasets. The synthesis phase presented two primary challenges: maintaining speaker identity and ensuring natural-sounding output. Despite using various TTS (Tacotron-Capacitron[31], GlowTTS[32]) and vocoder architectures (HiFi-GAN[33], UnivNet[34]), we encountered issues with mechanical articulation and high-frequency noise.

4.2.2. Evolution of Solutions

Our solution strategy evolved through three key phases:

1. **Audiobook Data Approach:** We first attempted using high-quality audiobook data (5-10 hours per speaker) from Amazon Audible. While this improved signal quality and reduced computational artifacts, particularly with HiFi-GAN vocoder, the synthetic speech exhibited notable monotonicity, lacking the dynamic range essential for political discourse.
2. **Enhanced Segmentation:** We implemented transcription-based sentence-level segmentation using Whisper Large Turbo's word-level timestamps as described in Section 3.2. This approach preserved linguistic coherence and improved phoneme alignment, leading to reduced noise and more accurate spectrogram generation.
3. **Advanced Model Architecture:** Finally, we transitioned to Few-Shot and Zero-Shot TTS models, which demonstrated superior performance compared to single-speaker training approaches, effectively addressing our remaining challenges.

This iterative progression from traditional approaches to more sophisticated solutions ultimately enabled us to achieve higher quality synthetic speech while maintaining speaker-specific characteristics and natural prosody.

5. Dataset Statistics and Analysis

To evaluate the perceptual quality and detection difficulty of synthetic speech in datasets, we conducted subjective and objective assessments. For subjective evaluation, we implemented a web-based listening test with 32 unique participants. Each participant was presented with 14 randomly selected audio samples (two from each dataset, one real and one fake) and tasked with classifying them as either genuine or synthetic speech. The results revealed significant variations in detection difficulty across datasets. Notably, our Famous Figures dataset achieved the highest mis-classification rate at 61.9%, suggesting that its synthetic speech samples more closely resemble natural speech compared to other datasets. The misclassification rates for different datasets are shown in Table 2 column Fake Miss-Rates.

For an objective quality assessment, we used the NISQA-TTS model [35], a deep learning-based model specifically designed to evaluate the quality of synthetic speech. As shown in Table 2, our Famous Figures dataset achieved impressive quality scores, with NISQA-TTS predicting naturalness of 3.69, surpassing both In the Wild and Spoof Celeb datasets.

6. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu,

"Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[3] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[4] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers, 2023," *URL: https://arxiv. org/abs/2301.02111. doi: doi*, vol. 10.

[5] J. Li and L. Zhang, "Zse-vits: A zero-shot expressive voice cloning method based on vits," *Electronics*, vol. 12, no. 4, p. 820, 2023.

[6] M. McLennan *et al.*, "Global Risks Report 2024." World Economic Forum, 2024. [Online]. Available: https://www.weforum.org/publications/global-risks-report-2024/digest/

[7] ebaker, "Russian War Report: Hacked news program and deepfake video spread false Zelenskyy claims," Mar. 2022.

[8] V. Elliott, "The Biden Deepfake Robocall Is Only the Beginning," *Wired*, Jan. 2024, section: tags. [Online]. Available: https://www.wired.com/story/biden-robocall-deepfake-danger/

[9] M. Spring, "Sadiq Khan says fake AI audio of him nearly led to serious disorder," Feb. 2024. [Online]. Available: https://www.bbc.com/news/uk-68146053

[10] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[11] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10155166/

[12] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[13] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 1–8.

[14] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[15] H. Ali, S. Subramani, S. Sudhir, R. Varahamurthy, and H. Malik, "Is audio spoof detection robust to laundering attacks?" in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024, pp. 283–288.

[16] J. Du, I.-M. Lin, I.-H. Chiu, X. Chen, H. Wu, W. Ren, Y. Tsao, H.-y. Lee, and J.-S. R. Jang, "Dfadd: The diffusion and flow-matching based audio deepfake dataset," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 921–928.

[17] Y. Xie, Y. Lu, R. Fu, Z. Wen, Z. Wang, J. Tao, X. Qi, X. Wang, Y. Liu, H. Cheng *et al.*, "The codecfake dataset and countermeasures for the universally detection of deepfake audio," *arXiv preprint arXiv:2405.04880*, 2024.

[18] 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," *International Joint Conference on Neural Networks (IJCNN)*, 2024.

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

[20] J.-w. Jung, Y. Wu, X. Wang, J.-H. Kim, S. Maiti, Y. Matsunaga, H.-j. Shim, J. Tian, N. Evans, J. S. Chung *et al.*, "Spoofceleb: Speech deepfake detection and sasv in the wild," *IEEE Open Journal of Signal Processing*, 2025.

[21] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[22] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "Xtts: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.

[24] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[25] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 682–689.

[26] S. Liao, Y. Wang, T. Li, Y. Cheng, R. Zhang, R. Zhou, and Y. Xing, "Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis," *arXiv preprint arXiv:2411.01156*, 2024.

[27] H. Wang, M. Yu, J. Hai, C. Chen, Y. Hu, R. Chen, N. Dehak, and D. Yu, "Ssr-speech: Towards stable, safe and robust zero-shot text-based speech editing and synthesis," *arXiv preprint arXiv:2409.07556*, 2024.

[28] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "Maskgct: Zero-shot text-to-speech with masked generative codec transformer," *arXiv preprint arXiv:2409.00750*, 2024.

[29] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2412.10117

[30] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. DAI, H. Lin, J. Chen, X. Du, L. Xue, Y. Chen, Z. Li, L. Xie, Q. Kong, Y. Guo, and W. Xue, "Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis," 2025. [Online]. Available: https://arxiv.org/abs/2502.04128

[31] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS," *preprint*, Apr. 2017. [Online]. Available: https://arxiv.org/pdf/1703.10135

[32] J. Kim, S. Kim, J. K. Zurbuchen, and S. Yoon*, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," *preprint*, Oct. 2020. [Online]. Available: https://arxiv.org/pdf/2005.11129

[33] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[34] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim1, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," *preprint*, Jun. 2021. [Online]. Available: https://arxiv.org/pdf/2106.07889

[35] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *arXiv preprint arXiv:2104.11673*, 2021.