# ArcticEcho: A Novel Speaker-Controlled Voice Cloning Dataset for Modern Deepfake Detection Benchmarking

Soham Gangopadhyay*, Inderpreet Singh†, Prateek Pandya‡, Ashish Mani§ and Sumit Goswami¶
* Amity University Uttar Pradesh, India. E-mail: sgangopadhyay2412@gmail.com
† Amity University Uttar Pradesh, India. E-mail: inderpreet1390@gmail.com
‡ Amity University Uttar Pradesh, India. E-mail: ppandya@amity.edu
§ Amity University Uttar Pradesh, India. E-mail: amani@amity.edu
¶ Defence Research and Development Organization (DRDO), India. Email: sumit.goswami@gov.in

***Abstract*—The rapid advancement of voice cloning technologies has fundamentally altered the landscape of audio deepfake detection, yet existing benchmarks have not evolved to reflect this reality. Current datasets inadvertently introduce confounding variables that enable detection models to exploit dataset-specific artifacts rather than learn genuine voice cloning signatures. To address this challenge, we introduce ArcticEcho, a novel speaker-controlled English language voice cloning dataset that eliminates confounding variables through rigorous experimental design. By maintaining strict correspondence between real and synthetic audio—identical speakers and identical content—ArcticEcho forces detection models to identify authentic voice cloning characteristics rather than incidental artifacts. Our dataset leverages state-of-the-art voice cloning technologies to create 24,752 high-quality audio samples across 18 speakers. Comprehensive evaluation on baseline models and a SOTA deep-learning NF-ResNeXt model reveals significant cross-dataset generalization gaps, with models experiencing substantial performance degradation when tested across different benchmarks. Our findings suggest that current benchmarking approaches may inadequately assess real-world deployment readiness, as detection systems achieving near-perfect performance on traditional datasets can struggle with high-quality synthetic speech. ArcticEcho provides a more realistic evaluation framework that better represents the challenges practitioners face with sophisticated voice cloning attacks, contributing to the development of more robust detection methods for modern audio security applications.**

## I. Introduction

The rapid advancement of voice cloning technologies has fundamentally altered the landscape of audio deepfake detection. While traditional text-to-speech systems produced synthetic speech with discernible artifacts, modern voice cloning methods generate high-fidelity audio that closely mimics human speech patterns, prosody, and speaker characteristics. This evolution presents a challenge: as synthetic speech quality improves, detection becomes more difficult, yet existing datasets have not evolved to reflect this reality.

Current audio deepfake detection datasets, including ASVSpoof [1]–[4] and FakeOrReal [5], were designed to simulate real-world conditions through diverse data collection. While valuable for advancing the field, this approach introduces confounding variables—differences in recording conditions, compression artifacts, speaker demographics, and content variations—that can allow detection models to learn spurious correlations rather than genuine deepfake signatures. Models may achieve high performance by exploiting quality-based heuristics or dataset-specific artifacts, leading to poor generalization when confronted with high-quality, modern voice cloning attacks.

The fundamental challenge lies in distinguishing between dataset artifacts and authentic voice cloning patterns. This limitation becomes critical as malicious actors increasingly employ sophisticated voice cloning tools that produce near-indistinguishable synthetic speech, necessitating detection methods that can identify genuine synthetic speech characteristics rather than incidental quality differences.

To address this challenge, we introduce ArcticEcho[1][6] , a novel speaker-controlled voice cloning dataset that eliminates confounding variables through rigorous experimental design. By maintaining strict correspondence between real and synthetic audio—identical speakers and identical content—ArcticEcho forces detection models to learn genuine voice cloning signatures rather than incidental artifacts. Our dataset leverages state-of-the-art voice cloning technologies (ElevenLabs [7], OpenVoice [8], Metavoice [9]) to create high-quality synthetic speech that reflects current threat landscapes.

We demonstrate that ArcticEcho's superior audio quality makes it more challenging for detection, not easier. This suggests that traditional datasets may inadvertently provide shortcuts to detection systems, enabling high performance without learning robust, generalizable mechanisms. Our cross-dataset validation experiments show significant performance degradation when models trained on existing datasets are evaluated on ArcticEcho, highlighting opportunities for improved benchmarking approaches.

ArcticEcho [6] offers a controlled approach to deepfake detection evaluation in English, complementing existing benchmarks while focusing on authentic voice cloning signatures. This precision in analysis could improve methods in practical

[1]ArcticEcho, 2025, Access at: https://doi.org/10.5281/zenodo.15838024

979-8-3315-7206-8/25/$31.00 

applications of identifying sophisticated synthetic speech.

## II. Related Works

The efficacy of audio deepfake detection systems hinges on the quality and design of their training and evaluation datasets. Benchmarks such as the ASVSpoof series [1], [3], [4], [10] and FakeOrReal [5] have significantly advanced research by providing standardized datasets for evaluating detection methods. However, these datasets often introduce confounding variables—such as channel variability, inconsistent recording conditions, and lack of controlled real-synthetic pairs—that can lead models to exploit dataset-specific biases rather than learning generalizable deepfake characteristics [11], [12]. ASVSpoof struggles to generalize to unknown attack types, with performance dropping significantly on unseen spoofing methods [1], [13]. Similarly, FakeOrReal suffers from data quality issues, including duplicate files and inconsistent bit rates, necessitating extensive preprocessing that may inadvertently bias detection models [14].

Concurrently, voice cloning technology has undergone a transformative evolution, shifting from traditional text-to-speech (TTS) systems to sophisticated voice cloning (VC) methods. Early TTS systems produced unnatural, robotic-sounding speech with limited speaker adaptation capabilities [15]. Modern VC approaches achieve significantly higher realism by incorporating advanced neural architectures and speaker adaptation techniques that can generate highly realistic synthetic voices from minimal reference audio, accurately capturing speaker-specific traits like tone, pitch, and prosody [16], [17]. This leap in realism poses a significant challenge for detection systems, as synthetic audio increasingly mirrors authentic human speech, rendering traditional detection methods less effective.

Traditional detection approaches, which often rely on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCC) or Constant-Q Cepstral Coefficients (CQCC) [14], [18], are increasingly inadequate in this evolving landscape. These feature-based methods are highly sensitive to variations in audio quality and can be easily circumvented by adversarial attacks, as demonstrated by [12]. Furthermore, their static nature limits adaptability to the growing sophistication of deepfake techniques, particularly those produced by state-of-the-art VC models [19]. This underscores the urgent need for detection strategies that are robust, adaptable, and capable of focusing on intrinsic deepfake artifacts rather than superficial dataset quirks.

To address these challenges, the research community has recognized the importance of controlled and diverse datasets. Efforts like the MLAAD dataset [13], which incorporates voice spoofs across multiple languages and TTS/VC models, represent progress toward broader evaluation benchmarks. However, such datasets still lack the stringent control over variables that is critical for isolating deepfake-specific features. ArcticEcho, introduced in this work, fills this gap by providing meticulously controlled real-cloned audio pairs, with exact speaker and transcript correspondence, generated using cutting-edge VC models. This design minimizes confounding variables, enabling detection methods to learn generalizable characteristics of audio deepfakes. By fostering the development of more robust and effective detection systems, ArcticEcho addresses the critical shortcomings of existing benchmarks and sets a new standard for audio deepfake research.

## III. Dataset Construction

### A. Source Data Selection

ArcticEcho is built upon the CMU Arctic speech corpus [20], a phonetically balanced dataset comprising 18 speakers with diverse acoustic characteristics. The corpus provides consistent recording conditions and speaker diversity across gender, age, and accent variations. Each speaker contributed approximately 1,150 utterances, totaling over 600 minutes of speech data.

### B. Voice Cloning Pipeline

We employed three state-of-the-art voice cloning systems representing different technological approaches:

- ElevenLabs [7] uses a three-stage pipeline: feature encoder extracts speaker embeddings, acoustic model generates mel spectrograms from text and speaker embeddings, and neural vocoder (likely HiFi-GAN-based) synthesizes the final waveform. The system supports both zero-shot Instant Voice Cloning and fine-tuned Professional Voice Cloning.
- OpenVoice [8] employs a decoupled architecture separating speech style from tone color. A base TTS model (modified VITS) generates initial speech with desired prosody, while a separate tone color converter using encoder-decoder with invertible normalizing flow transforms the generic voice to match reference speakers without retraining.
- Metavoice-1B [9] utilizes a multi-stage cascade with 1.2B parameters: EnCodec tokenizes audio into hierarchical representations, GPT-style causal transformer predicts initial token sequences, non-causal transformer adds fine acoustic details, and multi-band diffusion vocoder reconstructs the waveform with DeepFilterNet post-processing.

For each CMU Arctic utterance, we generated synthetic versions using the corresponding speaker's voice model and identical transcript. This speaker-content correspondence ensures detection performance differences arise from genuine voice cloning artifacts rather than content or speaker mismatches (Fig.1).

### C. Dataset Composition

The resulting ArcticEcho dataset contains **24,752 total files** comprising:

- **12,376 real utterances** from the original CMU Arctic corpus
- **12,376 synthetic utterances** generated through our voice cloning pipeline

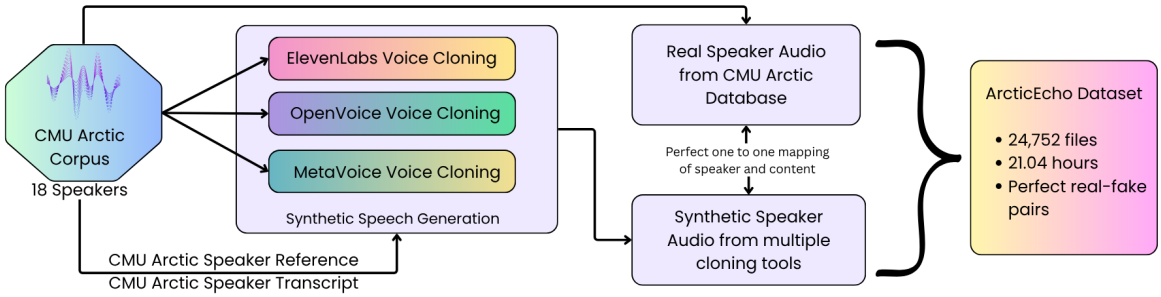


Fig. 1. The ArcticEcho Dataset Development Pipeline

TABLE I
GENERAL STATISTICS OF THE ARCTICECHO DATASET

| Metric | Value |
|---|---|
| Duration | 21.04 h (1262.4 min) |
| Avg. Length | 3.05 s |
| Total Files | 24,752 |
| Real/Fake Files | 12,376 / 12,376 |
| Speakers | 18 |

TABLE II
COMPARISON OF ARCTICECHO AND FOR DATASET NISQA QUALITY METRICS

| Metric | ArcticEcho | FoR | Improvement (%) | p-value |
|---|---|---|---|---|
| Overall (MOS) | 4.609 | 3.990 | 15.5 | 0.0000 |
| Noise | 4.138 | 3.638 | 13.8 | 0.0000 |
| Distortion | 4.565 | 4.324 | 5.6 | 0.0000 |
| Coloration | 4.288 | 4.008 | 7.0 | 0.0000 |
| Loudness | 4.420 | 4.171 | 6.0 | 0.0000 |

- **Total duration**: 21.04 hours with an average utterance length of 3.05 seconds
- **18 speakers** maintaining perfect speaker-content correspondence between real and synthetic pairs

All audio files are standardized to 16 kHz sampling rate with consistent encoding to eliminate technical artifacts as detection cues, ensuring models learn voice cloning signatures rather than encoding differences (Table I).

## IV. DATASET CHARACTERIZATION

### A. Audio Quality Analysis

To validate the superior quality of ArcticEcho compared to existing benchmarks, we conducted a comprehensive audio quality assessment using the NISQA (Non-Intrusive Speech Quality Assessment) framework [21]. The results demonstrate ArcticEcho's significant quality advantages across all perceptual dimensions.

ArcticEcho consistently outperforms FakeOrReal across all NISQA metrics, with statistically significant improvements ($p < 0.0001$) in overall quality, noise, distortion, coloration, and loudness measures (Table II).

TABLE III
CLASSIFICATION REPORT FOR REAL VS. FAKE CLASSIFICATION (OVERALL ACCURACY: 55.29%) BY HUMAN EVALUATORS

| Class | Precision (%) | Recall (%) | F1-Score (%) | Support |
|---|---|---|---|---|
| Real | 56.00 | 53.00 | 54.00 | 350 |
| Fake | 55.00 | 58.00 | 56.00 | 350 |

### B. Speaker Fidelity Verification

Speaker-similarity is examined through ECAPA-TDNN X-vectors, and detailed results are available in the Supplementary file. ArcticEcho demonstrates speaker good preservation with an average x-vector cosine similarity of $0.776 \pm 0.063$ across 12,376 real-fake pairs, indicating moderate maintenance of perceptual speaker identity while introducing synthetic artifacts that detection systems must identify. Speaker-level analysis shows performance variation across all 18 speakers (similarities 0.6598–0.8882), showing the challenge of voice cloning diverse vocal characteristics.

To evaluate human perception of our synthetic speech, we conducted a listening test with 35 evaluators (forensic science doctoral and graduate students under faculty supervision) on 20 randomly selected audio files (10 real, 10 fake). Human evaluators achieved only 55.29% overall accuracy (Table III), with balanced performance on both real and fake samples. The detailed human evaluator performance data is available in the Supplementary file. This near-chance performance confirms the high perceptual quality of our synthetic speech, validating ArcticEcho's effectiveness as a challenging benchmark for both human and automated detection systems.

### C. Prosodic Feature Preservation

Modern voice cloning technology effectively preserves prosodic features, which we measure across 17 comprehensive measures including F0, energy, spectral, temporal, voice quality, and formant characteristics. Overall statistics of feature preservation across the dataset are shown in Table IV.

## V. EXPERIMENTAL VALIDATION

To evaluate ArcticEcho as a benchmark for audio deepfake detection, we conducted comprehensive experiments using traditional feature extraction methods and machine learning

TABLE IV
SPEAKER SIMILARITY ANALYSIS (BETWEEN REAL AND FAKE PAIRS): PROSODIC FEATURES AND X-VECTOR EMBEDDINGS

| Feature | Similarity | Feature | Similarity | Feature | Similarity |
|---|---|---|---|---|---|
| X-vector | $0.776 \pm 0.063$ | F0 mean | $0.973 \pm 0.014$ | RMS mean | $0.979 \pm 0.017$ |
| Spec. centroid mean | $0.972 \pm 0.024$ | ZCR mean | $0.886 \pm 0.057$ | Tempo | $0.979 \pm 0.014$ |
| Jitter | $0.924 \pm 0.061$ | Shimmer | $0.943 \pm 0.036$ | HNR | $0.944 \pm 0.040$ |
| F2 formant | $0.970 \pm 0.017$ | Prosodic features avg. | $0.940 \pm 0.037$ | – | – |

classifiers. Our evaluation framework focuses on within-dataset performance and cross-dataset generalization capabilities.

### A. Feature Extraction Methods

We employed three widely-used feature extraction methods representing different approaches to capturing audio characteristics. Mel-Frequency Cepstral Coefficients (MFCC) serve as traditional spectral features that capture the general shape of the spectral envelope. We extracted 13 MFCC coefficients along with their first and second derivatives, resulting in 39-dimensional feature vectors. Linear-Frequency Cepstral Coefficients (LFCC) provide an alternative to MFCC using linear frequency scaling instead of mel-scaling, potentially better suited for detecting synthetic speech artifacts. We extracted 13 LFCC coefficients with derivatives, yielding 39-dimensional features. Constant-Q Cepstral Coefficients (CQCC) represent advanced spectral features designed specifically for anti-spoofing applications, utilizing constant-Q transform to provide better frequency resolution at lower frequencies. We extracted 19 CQCC coefficients with derivatives, resulting in 57-dimensional features.

### B. Classification Methods

We implemented three complementary machine learning approaches to provide comprehensive evaluation coverage. Logistic Regression (LR) serves as a linear classifier that establishes baseline performance and tests whether features are linearly separable. Support Vector Machine (SVM) with RBF kernel functions as a non-linear classifier that can capture complex decision boundaries. Random Forest (RF) operates as an ensemble method that provides robustness and feature importance insights while handling non-linear relationships. We also report performance evaluation of the ArcticEcho dataset on a SOTA NF-ResNeXt [22] deep learning model.

### C. Evaluation Protocol

Our evaluation protocol consisted of two primary components designed to assess both internal consistency and external generalizability. For within-dataset evaluation, we evaluated all feature-model combinations using Leave-One-Speaker-Out (LOSO) cross-validation on ArcticEcho to ensure robust performance estimates and prevent speaker-specific overfitting. For cross-dataset evaluation, we conducted bidirectional cross-dataset validation between ArcticEcho and FakeOrReal, training on one dataset and testing on the other in both directions. Our analysis framework computed generalization gaps as the difference between within-dataset and cross-dataset performance to quantify dataset-specific biases and experimental control.

## VI. RESULTS AND ANALYSIS

### A. Within-Dataset Performance Analysis

The within-dataset results reveal important differences in detection challenge levels across benchmarks. ArcticEcho demonstrates more realistic performance ranges (0.760-0.925) compared to FakeOrReal's near-perfect accuracies (0.929-0.999), as shown in Table V. On the SOTA NF-ResNeXt model, ArcticEcho achieves an EER of 7.25%, as compared to EER of 5.13% on the ASVSpoof 2019 LA dataset, showing a relative degradation of 29.24%. This difference suggests that ArcticEcho better captures the genuine difficulty of modern voice cloning detection, where even sophisticated models face meaningful challenges.

LFCC features consistently outperform MFCC and CQCC across all classifiers on ArcticEcho, achieving the highest accuracy of 0.925 (SVM). This suggests that linear frequency scaling may be better suited for capturing the spectral characteristics introduced by modern voice cloning algorithms. CQCC features, while showing more modest peak performance, demonstrate remarkable consistency across different classifiers, indicating more stable detection characteristics.

### B. Cross-Dataset Generalization Insights

The cross-dataset evaluation reveals substantial generalization challenges that highlight important practical considerations. Models trained on ArcticEcho achieve 0.508-0.721 accuracy on FakeOrReal, while FakeOrReal-trained models achieve 0.529-0.656 accuracy on ArcticEcho. These performance drops suggest that both datasets capture different aspects of the detection challenge, with important implications for real-world deployment.

The feature robustness rankings from cross-dataset evaluation (Table VI) reveal a critical insight: CQCC features demonstrate the highest cross-dataset performance (0.690) despite modest within-dataset performance. This transferability versus peak performance trade-off suggests that features optimized for specific conditions may sacrifice generalizability—an important consideration for practical deployment where data distribution shifts are common.

### C. Speaker-Level Vulnerability Patterns

Individual speaker analysis reveals consistent vulnerability patterns across all feature sets (Table VII). Speaker 'slp'

TABLE V
CORE ACCURACY METRICS FOR CROSS-DATASET ANALYSIS BETWEEN ARCTIC AND FOR DATASETS.

| Feature Set | Model | Within Arctic | Within FoR | Arctic→FoR | FoR→Arctic | Generalization Gap |
|---|---|---|---|---|---|---|
| MFCC | LR | 0.793 | 0.933 | 0.508 | 0.541 | 0.251 |
| | SVM | 0.855 | 0.999 | 0.596 | 0.542 | 0.313 |
| | RF | 0.778 | 0.995 | 0.553 | 0.529 | 0.249 |
| LFCC | LR | 0.846 | 0.985 | 0.646 | 0.608 | 0.239 |
| | SVM | 0.925 | 0.999 | 0.652 | 0.553 | 0.372 |
| | RF | 0.813 | 0.997 | 0.642 | 0.575 | 0.239 |
| CQCC | LR | 0.809 | 0.929 | 0.641 | 0.644 | 0.165 |
| | SVM | 0.808 | 0.993 | 0.710 | 0.660 | 0.148 |
| | RF | 0.760 | 0.980 | 0.721 | 0.656 | 0.104 |

TABLE VI
FEATURE AND MODEL ROBUSTNESS RANKINGS (ACCURACY)

| Feature Robustness | | | Model Robustness | | |
|---|---|---|---|---|---|
| Rank | Feature | Avg Cross-Perf. | Rank | Model | Avg Cross-Perf. |
| 1 | CQCC | 0.690 | 1 | SVM | 0.652 |
| 2 | LFCC | 0.647 | 2 | RF | 0.638 |
| 3 | MFCC | 0.552 | 3 | LR | 0.598 |

TABLE VII
DETECTION DIFFICULTY ANALYSIS (SPEAKER-WISE)

| Feature Set | Hardest Speakers | | Easiest Speakers | |
|---|---|---|---|---|
| | Speaker | Avg Accuracy | Speaker | Avg Accuracy |
| MFCC | awb | 0.448 | gka | 0.912 |
| | slp | 0.531 | rms | 0.935 |
| | clb | 0.641 | aew | 0.956 |
| | ksp | 0.743 | ljm | 0.956 |
| | fem | 0.775 | rxr | 0.956 |
| LFCC | slp | 0.630 | ksp | 0.907 |
| | axb | 0.743 | awb | 0.919 |
| | fem | 0.782 | rxr | 0.926 |
| | slt | 0.783 | ljm | 0.950 |
| | ahw | 0.809 | aew | 0.981 |
| CQCC | slp | 0.541 | axb | 0.944 |
| | aup | 0.606 | ksp | 0.953 |
| | awb | 0.620 | rxr | 0.961 |
| | jmk | 0.628 | slt | 0.966 |
| | gka | 0.672 | rms | 0.972 |

consistently emerges as the most challenging for detection (0.541-0.630 accuracy), while speakers 'rxr' and 'aew' are consistently the easiest (0.926-0.981 accuracy). This consistency across different feature extraction methods suggests intrinsic voice characteristics that influence voice cloning vulnerability. Analysis reveals accent-dependent difficulty patterns, with non-American accents showing increased detection challenges.

These patterns indicate that certain vocal characteristics make individuals inherently more susceptible to high-quality voice cloning attacks. This finding has practical implications for personalized security measures and risk assessment frameworks, providing insights that would be impossible to discover in uncontrolled datasets where such patterns are confounded by other variables.

### D. Feature Robustness and Practical Implications

The substantial performance gaps between within-dataset and cross-dataset evaluation (averaging 0.231) reveal important considerations for practical deployment. The fact that models achieving near-perfect performance on one dataset can experience dramatic degradation on another suggests potential deployment readiness challenges when detection systems encounter data distributions different from their training conditions.

MFCC features show the largest generalization gaps (0.249-0.313), while CQCC features demonstrate the smallest (0.104-0.165). This suggests that different features have different robustness profiles to distribution shifts, informing feature selection strategies for practical applications where robustness may be more important than peak performance.

### E. Implications for Real-World Deployment

The performance patterns observed in our evaluation highlight important considerations for real-world deployment of deepfake detection systems. The modest performance levels achieved on ArcticEcho (80-90% range) may be more representative of what practitioners face with high-quality attacks, compared to the near-perfect accuracies often achieved on traditional benchmarks.

The quality-difficulty relationship revealed by ArcticEcho suggests that as voice cloning technology advances, detection systems must be prepared for increasingly sophisticated attacks that maintain high perceptual quality while introducing subtle synthetic artifacts. Evolving threats require detection methods that identify authentic voice cloning signatures, not quality-based shortcuts.

ArcticEcho's controlled design enables precise analysis of detection system capabilities, potentially informing robust methods for high-quality voice cloning threats. These insights contribute to the enhanced understanding of practical challenges in audio deepfake detection.

## VII. CONCLUSION

In this work, we introduced ArcticEcho, a novel speaker-controlled voice cloning dataset that eliminates confounding variables present in existing datasets through strict speaker-content correspondence. Our controlled approach enables more precise analysis of voice cloning detection mechanisms.

The core contribution of this work is the ArcticEcho dataset itself. Unlike existing resources, ArcticEcho provides clean separation of speaker identity and linguistic content, ensuring that detection models cannot exploit spurious correlations or quality-based shortcuts. This enables rigorous benchmarking of detection systems under conditions that better reflect modern, high-quality synthetic speech.

The speaker-level analysis reveals consistent vulnerability patterns, indicating certain vocal characteristics influence voice cloning susceptibility—insights impossible to discover in uncontrolled datasets. Additionally, feature robustness profiles show transferability and peak performance trade-offs, informing practical feature selection strategies.

We make ArcticEcho [6] publicly available to support development of more robust detection methods suited for modern voice cloning threats. The controlled framework provides a foundation for systematic evaluation and enables better understanding of practical challenges facing audio deepfake detection systems. ArcticEcho focuses on English speech, providing a controlled framework for systematic evaluation with future multilingual extensions planned.

## REFERENCES


[1] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, ISCA: ISCA, Sep. 2019.

[2] X. Wang, H. Delgado, H. Tak, *et al.*, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," 2024. eprint: 2408.08739 (eess.AS).

[3] X. Liu, X. Wang, M. Sahidullah, *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2507–2522, 2023.

[4] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, ISCA: ISCA, Sep. 2021.

[5] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Timisoara, Romania: IEEE, Oct. 2019.

[6] S. Gangopadhyay and I. Singh, *ArcticEcho: A Speaker-Controlled voice cloning dataset for modern deepfake detection benchmarking*, 2025.

[7] ElevenLabs, *Voice cloning technology*, Accessed: 2025-07-08, 2025. [Online]. Available: https://elevenlabs.io/voice-cloning.

[8] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile instant voice cloning," 2023. eprint: 2312.01479 (cs.SD).

[9] MetaVoice, *Metavoice-1b: A multilingual voice cloning model*, Accessed: 2025-07-08, 2024. [Online]. Available: https://github.com/metavoiceio/metavoice-src.

[10] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," en, *Comput. Speech Lang.*, vol. 64, no. 101114, p. 101 114, Nov. 2020.

[11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, *et al.*, "Shortcut learning in deep neural networks," en, *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[12] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," 2022. eprint: 2203.16263 (cs.SD).

[13] N. M. Müller, P. Kawa, W. H. Choong, *et al.*, "MLAAD: The Multi-Language audio Anti-Spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*, vol. 33, Yokohama, Japan: IEEE, Jun. 2024, pp. 1–7.

[14] A. Hamza, A. R. R. Javed, F. Iqbal, *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[15] H. Azzuni and A. E. Saddik, "Voice cloning: Comprehensive survey," 2025. eprint: 2505.00579 (cs.SD).

[16] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/4559912e7a94a9c32b09d894f2bc3c82-Paper.pdf.

[17] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020.

[18] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio deepfake detection: What has been achieved and what lies ahead," en, *Sensors (Basel)*, vol. 25, no. 7, p. 1989, Mar. 2025.

[19] O. A. Shaaban and R. Yildirim, "Audio deepfake detection using deep learning," en, *Eng. Rep.*, vol. 7, no. 3, Mar. 2025.

[20] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech 2021*, ISCA: ISCA, Aug. 2021.

[22] Z. Zhang, X. Zhao, and X. Yi, "Improving robustness of speech anti-spoofing system using resnext with neighbor filters," *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251846550.