Evaluating speaker verification robustness to neural voice cloning attacks*

Nazarii Dzhaliuk[1,*,†], Khrystyna Ruda[1,†], Dmytro Sabodashko[1,†], and Volodymyr Khoma[1,†]

[1] Lviv Polytechnic National University, Information Security Department, 12 Stepan Bandera str., 79000 Lviv, Ukraine

Abstract

Recent advances in neural voice cloning have increased the susceptibility of speaker verification systems to synthetic speech attacks. This paper evaluates the robustness of three detection paradigms: classical machine-learning methods with handcrafted features, an end-to-end raw-waveform model (RawNet2), and hybrid Vision Transformer (ViT) architectures applied to time–frequency representations. A custom dataset of 3,600 audio samples generated using XTTS, Tortoise, ElevenLabs, and RVC was used under text-dependent and text-independent conditions. Experimental results show that ViT-based models, particularly with MFCC inputs, achieve the best overall performance (70.5% accuracy, F1-score 0.727), outperforming RawNet2 and classical approaches. However, vendor-specific evaluation reveals a significant generalization gap, with detection accuracy dropping to near-random levels for commercial black-box systems. These results highlight the need for more diverse training data to improve robustness against modern voice cloning technologies.

Keywords

voice cloning detection, deepfake audio, vision transformers, RawNet2, ElevenLabs, cybersecurity.

1. Introduction

With the rapid development of artificial intelligence (AI) technologies and the expansion of their application areas, new challenges for cybersecurity have emerged, characterized by novel vulnerabilities and creative utilizations of AI models to exploit existing security gaps. A prominent application of this evolution is speech synthesis. While possessing immense potential for legitimate sectors, ranging from business applications like voice assistants and media production to accessibility tools for individuals with medical conditions, speech synthesis has inevitably fostered the parallel development of sophisticated voice cloning technologies.

The efficiency of these systems has increased dramatically; technological advancements have reduced the required audio sample duration for effective cloning from minutes to mere seconds. This rapid progression outpaces the development of detection mechanisms, creating a new vector of threats in the field of information security. Consequently, the identification of synthetic voices has become a critical objective. This paper addresses this necessity by analyzing existing approaches to voice cloning detection and evaluating their individual effectiveness, specifically focusing on the comparative analysis of neural network-based methods and classical machine learning techniques.

The Escalation of AI-Driven Social Engineering and Financial Fraud

The proliferation of voice cloning technologies has fundamentally altered the threat landscape of social engineering and financial fraud. Driven by rapid advancements in deep learning, the global market for voice cloning is projected to reach approximately $3.1 billion by 2030th. This economic

* CPITS 2026: Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2026, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

nazarii.o.dzhaliuk@lpnu.ua (N. Dzhaliuk); khrystyna.s.ruda@lpnu.ua (K. Ruda); dmytro.v.sabodashko@lpnu.ua (D. Sabodashko); volodymyr.v.khoma@lpnu.ua (V. Khoma)

0009-0006-9381-416X (N. Dzhaliuk); 0000-0001-8644-411X (K. Ruda); 0000-0003-1675-0976 (D. Sabodashko); 0000-0001-9391-6525 (V. Khoma)

growth runs parallel to a dramatic increase in the efficiency of synthesis models; notably, systems such as Baidu's Deep Voice reduced the required reference audio duration for effective cloning from 30 minutes to merely 3.7 seconds within a single year [1].

The accessibility of these tools has lowered the barrier to entry for cybercriminals, leading to a surge in sophisticated attacks [3]. Financial institutions and the insurance sector are particularly vulnerable; a 2024 report by Zscaler ThreatLabz indicated a 393% increase in phishing attacks, with AI-driven voice cloning identified as a primary trend [4]. Attack vectors have evolved from generic phishing to complex “vishing” (voice phishing) scenarios, where attackers impersonate clients to bypass verification protocols or simulate the voices of family members in distress to solicit funds. Furthermore, the technology is increasingly utilized in disinformation campaigns targeting public figures, journalists, and politicians, thereby amplifying reputational risks and societal instability. The widespread availability of unregulated commercial and open-source synthesis platforms further exacerbates these risks, enabling the rapid deployment of high-quality deepfakes without robust identity verification.

Vulnerabilities in Current Voice Biometric Systems

The evolution of voice cloning poses a critical threat to Automatic Speaker Verification (ASV) systems, which are increasingly deployed for biometric authentication in banking and remote services. These systems are highly susceptible to spoofing attacks, where synthesized or converted speech is used to impersonate a legitimate user. To mitigate this risk, modern biometric architectures typically incorporate Presentation Attack Detection (PAD) systems. Conceptually, the PAD module functions as a preliminary gatekeeper: it analyzes the input utterance for artifacts of synthesis and rejects the request if spoofing is detected, thereby preventing the signal from reaching the ASV module for identity verification.

However, the efficacy of current PAD solutions remains inconsistent. Research indicates significant vulnerabilities when adversarial transformations are applied to the audio signal. For instance, simple post-processing techniques, such as the addition of background noise or background music, can successfully obfuscate the artifacts generated by synthesis models [2], allowing spoofed audio to bypass detection mechanisms. Furthermore, a critical limitation of existing PAD systems is their lack of generalization. While they may effectively detect attacks from known synthesis algorithms (seen during training), they often fail against novel, high-fidelity neural synthesizers. This "arms race" dynamic, where the speed of synthesis development outpaces detection capabilities, creates a significant security gap, necessitating the development of more robust, architecture-agnostic detection methods.

To address the generalization gap identified in current PAD solutions, this study presents a comprehensive comparative analysis of three distinct detection paradigms: Classical Machine Learning (utilizing Support Vector Machines with manual feature extraction), Hybrid architectures (Vision Transformers applied to MFCC, Mel-Spectrogram, and CQT representations), and End-to-End Deep Learning (RawNet2). The primary contribution of this work is the rigorous evaluation of these architectures against a diverse threat landscape, utilizing a custom-curated dataset of 3,600 audio samples.

Unlike previous studies that often rely on homogeneous datasets, this research evaluates robustness against four specific modern generator types representing different synthesis techniques: autoregressive models (Tortoise), zero-shot systems (XTTS), retrieval-based voice conversion (RVC), and proprietary commercial engines (ElevenLabs). Furthermore, the study introduces a methodological distinction between text-dependent (ST) and text-independent (DT) scenarios to assess whether detectors rely on acoustic artifacts or semantic patterns. Finally, beyond detection accuracy, the research evaluates the practical feasibility of these models for edge deployment by quantifying inference time and memory consumption, thereby providing a holistic view of the trade-offs between architectural complexity and operational efficiency.

2. Analysis of Recent Research and Publications

To establish a comprehensive detection framework, this study synthesizes findings from 33 key literary sources and evaluates the architectural constraints of ten open-source detection implementations initially identified during the preliminary analysis. We categorize the evolving threat landscape into four primary generative paradigms that currently dominate the voice cloning market: autoregressive, non-autoregressive, diffusion-based, and retrieval-based voice conversion systems.

The rapid proliferation of artificial intelligence has fundamentally reshaped the cybersecurity landscape, introducing novel attack vectors that leverage machine learning to exploit biometric vulnerabilities. In the domain of audio synthesis, this evolution is characterized by the transition from traditional statistical methods to neural voice cloning systems capable of generating high-fidelity speech with minimal reference data. The current market offers a diverse array of synthesis tools, ranging from open-source repositories to proprietary commercial platforms, which significantly lowers the barrier to entry for malicious actors regarding social engineering and spoofing attacks.

Consequently, the development of robust detection mechanisms requires a comprehensive understanding of the underlying architectures of these synthesis engines. Because the field involves a constant "arms race" between generation and detection, defensive models must adapt to the specific artifacts introduced by different generative paradigms. This section reviews the evolution of neural voice cloning technologies, analyzes the specific synthesis techniques evaluated in this study-including autoregressive and diffusion-based models-and discusses the application of artificial intelligence in defensive cybersecurity strategies.

Modern neural voice cloning systems utilize distinct deep learning architectures, each presenting unique characteristics regarding synthesis quality and computational efficiency. A prevalent approach involves autoregressive models, which generate audio sequences step-by-step, predicting the current element based on previous spectrogram frames or audio fragments. This sequential dependency allows for high-fidelity reproduction of intonation and timbre, making these models highly effective for expressive speech synthesis. However, the iterative nature of autoregressive generation results in significant latency and potential instability when generating long audio sequences, where error accumulation can degrade quality [5].

In contrast, non-autoregressive models generate the entire audio sequence in parallel. This parallelism significantly reduces inference time compared to autoregressive counterparts and improves stability by preventing the propagation of generation errors [5]. While these models offer efficiency, they often produce “flatter” prosody, lacking the nuanced emotional and rhythmic variations found in natural human speech, as the global generation approach may smooth out fine-grained acoustic details [6, 7].

A more recent advancement is the application of diffusion probabilistic models. These models operate by gradually removing noise from a signal; during training, they learn to reverse a diffusion process that adds Gaussian noise to data, eventually recovering a clean waveform from pure noise (Figure 1). Diffusion models excel at capturing complex acoustic distributions, resulting in highly natural timbre and emotional subtlety. However, this iterative "denoising" process requires hundreds of steps, making diffusion-based synthesis computationally expensive and slow compared to other paradigms [5].

This study specifically investigates architectures that facilitate rapid voice adaptation. Zero-shot voice cloning, exemplified by the XTTS architecture [8], represents a significant shift from traditional model fine-tuning [9]. XTTS is a multilingual end-to-end model based on the Tortoise architecture but optimized for speed [10]. Unlike systems requiring extensive retraining for a target speaker, zero-shot models utilize trained encoders to extract a vector representation (embedding) of

a speaker's voice characteristics from a short reference audio sample (often just a few seconds). This embedding conditions the synthesis engine to generate new speech in the target voice immediately. While efficient, zero-shot approaches relying on generalization may lack the granular detail and specific stylistic quirks captured by models trained extensively on a single speaker [11–16].

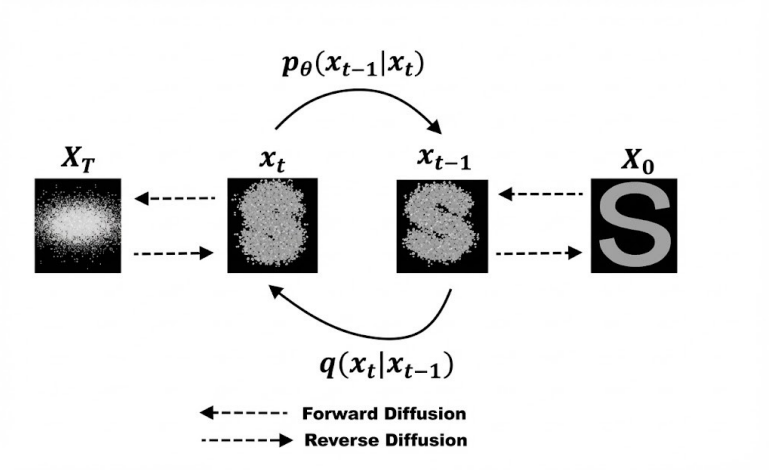

Figure 1: Architecture of the diffusion model

Alternatively, Retrieval-based Voice Conversion (RVC) operates on a different principle: rather than synthesizing speech from text “from scratch,” it converts an existing audio signal into the target voice. The RVC model separates the source audio's content (context) from its style. During inference, the model retrieves features from the target speaker's dataset to reconstruct the source content in the target style. This approach allows for extremely high-fidelity results and can function in near real-time, though it requires a “calibration” phase where the model familiarizes itself with the target voice’s specific features.

A critical challenge in forensic audio analysis is the distinction between open-source and proprietary commercial synthesis engines. Open-source models, such as Tortoise and RVC, have transparent architectures and publicly available implementations [8, 10], allowing researchers to study their artifacts and train specific detectors. Tortoise, for instance, combines autoregressive Transformers for spectrogram generation with diffusion decoders, creating a complex but analyzable artifact footprint.

Conversely, commercial platforms like ElevenLabs operate as "black box" systems. ElevenLabs offers advanced services including text-to-speech, voice cloning, and dubbing [17], often achieving state-of-the-art naturalness that rivals human speech. However, because the underlying architecture is proprietary and closed-source, little is publicly known about its specific internal mechanisms, aside from general claims regarding safety monitoring and watermarking. This lack of architectural transparency creates a "generalization gap" for detection systems; detectors trained on known, open-source algorithms often fail to generalize to the sophisticated, unknown artifacts produced by optimized commercial engines, creating a significant vulnerability in current security frameworks.

3. Methodology and System Architectures

Traditionally, network access is described using fields such as IP addresses, protocol, port, and actions allow or deny. This approach to regulating network access operates primarily at Layers 3 and 4 of the OSI model and, in practice, grants access not to a specific host or system, but rather to

whoever sends packets with matching values in the “source IP,” “destination IP,” and “protocol/port” fields. Early firewall technologies were based on these principles, which was largely acceptable at the time. However, organizing network access control in accordance with Zero Trust principles using only basic conditions (IP and port fields) is insufficient to ensure the security of modern network architectures due to limited flexibility, excessive static behavior, lack of sender authentication, and restricted visibility.

3.1. Dataset Construction

A critical limitation in existing voice cloning research is the lack of datasets that adequately represent the "black box" nature of modern commercial synthesizers. To overcome this, a custom dataset was curated comprising 3,600 audio samples derived from 20 unique speakers. The dataset is strictly partitioned into real and synthetic subsets to ensure objective evaluation.

- Real Samples: The control group consists of 400 authentic recordings, collected from open sources, representing 20 distinct speakers with 20 samples per speaker.
- Synthetic Samples: The experimental group contains 3,200 cloned samples generated using four distinct synthesis engines: ElevenLabs, RVC, XTTS, and Tortoise.

A novel methodological contribution of this study is the segmentation of synthetic data into two distinct testing scenarios to evaluate semantic overfitting (Figure 2):

1. Same Text (ST): A subset of 1,600 samples where the cloned voice articulates the exact text spoken in the original reference audio. This scenario tests the system's ability to detect acoustic artifacts without semantic discrepancies.
2. Different Text (DT): A subset of 1,600 samples where the cloned voice articulates text different from the original reference. This scenario simulates realistic social engineering attacks where the attacker generates novel content.

The dataset encompasses various audio formats (.wav, .mp3) and sampling rates (20kHz, 24kHz, 48kHz), with a total duration of approximately 7 hours and 15 minutes for the synthetic class and 55 minutes for the real class.

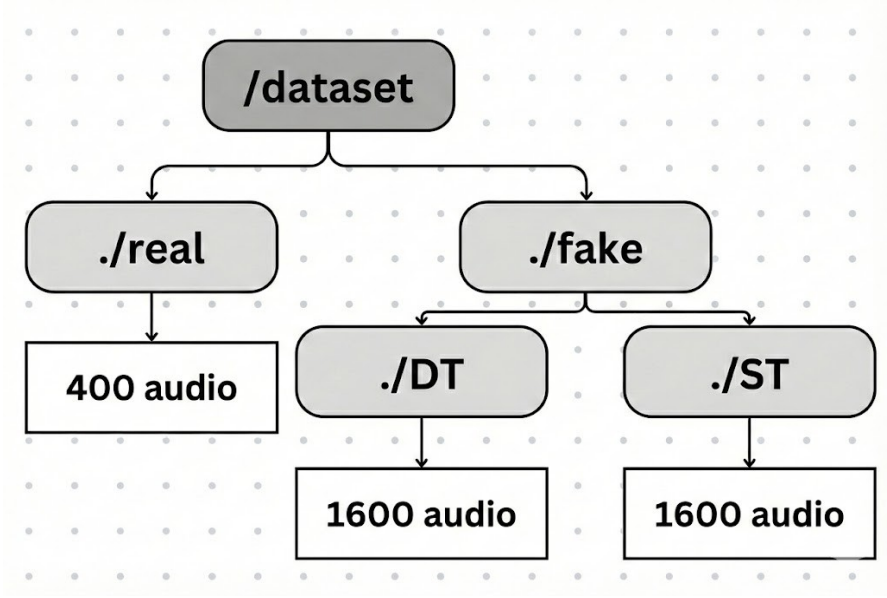

Figure 2: Dataset structure

3.2. Hybrid approach: Vision Transformers (ViT)

The hybrid approach treats audio forgery detection as a computer vision classification task. This architecture leverages the Vision Transformer (ViT) model, originally designed for image recognition, to analyze visual representations of audio signals (Figure 3) [18].

Feature Extraction (Visual Acoustic Representations) Unlike end-to-end models, the hybrid system requires a preprocessing stage to convert 1D audio waveforms into 2D spectrograms. Three distinct feature extraction techniques were evaluated to determine the optimal input representation for the Transformer:

- MFCC (Mel-Frequency Cepstral Coefficients): Describes the power spectral envelope of a signal derived via the Discrete Cosine Transform (DCT) [19] of log-powers on the Mel scale. This method provides a compact representation of the signal's energy distribution.
- Mel-Spectrogram: A visual representation of the signal's frequency spectrum over time, mapped to the Mel scale. Unlike MFCCs, this retains more granular data, providing a richer "image" for the Convolutional Neural Networks (CNN) or Transformers to analyze.
- Constant-Q Transform (CQT): A spectral transform utilizing a logarithmic frequency scale with a constant Q-factor (ratio of center frequency to bandwidth) [20]. This representation is particularly effective for analyzing tonal structures and harmonic distortions often present in synthetic speech.

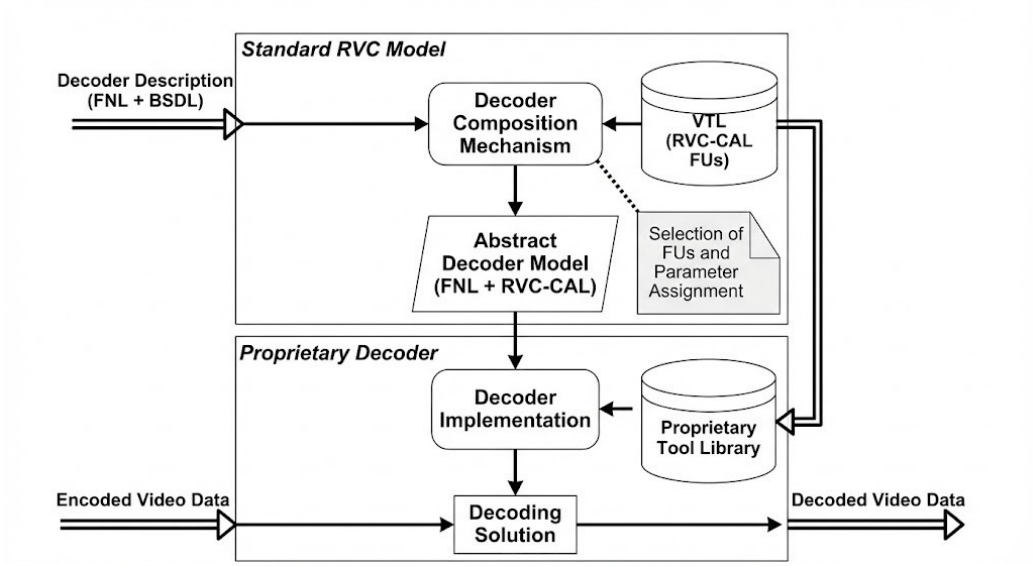

Figure 3: Architecture of the RVC model

The generated spectrograms are processed by the ViT model, which divides the image into non-overlapping patches of N×N pixels. These patches are flattened into vectors and augmented with Positional Encodings (PE) to retain temporal and spectral order. The architecture utilizes a special Classification Token (CLS) to aggregate information across the sequence via Multi-Head Self-Attention (MSA) mechanisms. The network employs Multi-Layer Perceptron (MLP) blocks with Gaussian Error Linear Unit (GELU) activation functions to capture non-linear dependencies within the acoustic features.

3.3. End-to-end approach: RawNet2

To evaluate the efficacy of feature learning from raw data, we implemented the RawNet2 architecture [21]. This end-to-end approach bypasses traditional hand-crafted feature extraction (such as STFT), operating directly on the raw waveforms.

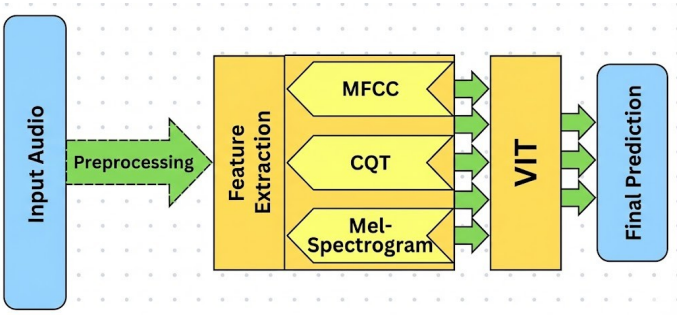

Figure 4: Architecture of ViT model

The architecture comprises two primary modules:

1. Front-end (Embedding Extraction): This module utilizes a modified CNN-GRU architecture. It processes the raw input using ResNet blocks-deep convolutional networks with residual connections-to extract local features. These features are subsequently processed by Gated Recurrent Units (GRU), a streamlined variant of LSTM (Long Short-Term Memory) networks, to model temporal dependencies without the vanishing gradient problems associated with standard Recurrent Neural Networks (RNNs) [21].
2. Back-end (Classifier): The extracted speaker embeddings are processed using a "concat & mul" method, where feature vectors are concatenated and element-wise multiplied before being passed to a deep fully connected network for binary classification [22].

The model employs Leaky ReLU activation functions to maintain gradient flow and prevent "dead neurons" during training. This architecture is theoretically capable of detecting subtle phase information and waveform artifacts that might be lost during the conversion to spectrograms in hybrid models.

3.4. Baseline classical approach

To establish a performance baseline and verify the necessity of Deep Learning, a classical machine learning approach was implemented. This method relies on manual feature engineering followed by statistical classification.

The system extracts musical and harmonic features rather than purely spectral ones, utilizing:

- Chroma Features (Chromagram): A 12-element vector representing the energy distribution across pitch classes (C, C#, D, etc.). This feature captures harmonic structures and is robust to variations in tone, potentially identifying the "perfect" harmonic alignment often generated by synthesis algorithms.
- Tonnetz: Represents tonal centroids to capture harmonic relationships between notes, providing data on the emotional and harmonic stability of the speech.

The extracted features are classified using a Support Vector Machine (SVM) operating in a frame-based multiclass configuration. The SVM evaluates individual audio frames independently, aggregating scores to determine the most probable class (real or fake). This approach is computationally lightweight but lacks the capacity to model complex, non-linear abstractions or long-term temporal dependencies inherent in high-fidelity deepfakes.

4. Experimental Results and Analysis

This section presents the empirical evaluation of the implemented detection systems. The assessment was conducted in two phases: a performance evaluation focusing on classification metrics (Accuracy, F1-score) and a computational efficiency analysis measuring inference time and

memory consumption. All experiments utilized the custom dataset described in Section 3.1, comprising 3,600 audio samples partitioned into "Same Text" (ST) and "Different Text" (DT) scenarios to isolate semantic bias.

The comparative analysis of the three architectural paradigms reveals a distinct hierarchy in detection capabilities. The Hybrid ViT architecture [23] utilizing MFCC features achieved the highest overall performance, recording an accuracy of 0.705 and an F1-score of 0.727 on the text-independent (DT) test set. Evaluation metrics were calculated using the scikit-learn module for Python [23]. This suggests that the compact energy distribution representation provided by MFCCs is currently the most effective feature set for Transformer-based detection among those tested.

The End-to-end RawNet2 model demonstrated competitive performance with an accuracy of 0.659 and an F1-score of 0.663. While slightly trailing the ViT-MFCC model, RawNet2 proved significantly more robust than the baseline approaches, validating the efficacy of learning directly from raw waveforms without manual feature extraction.

In stark contrast, the classical machine learning baseline (SVM with manual feature engineering) failed to distinguish modern deepfakes effectively. The classical models yielded accuracies as low as 0.232 (csun22) [24, 25] and 0.233 (sudarshanng7), with F1-scores dropping below 0.150. This fundamental failure underscores that traditional features, such as harmonic alignment or simple spectral statistics, are insufficient for detecting the sophisticated artifacts generated by neural synthesis engines.

Table 1
Comparative performance metrics of implemented architectures on the DT (Different Text) dataset

Architecture Type	Model Implementation	Accuracy (DT)	F1-Score	Inference Time (s)
Hybrid ViT	ViT-MFCC (Best)	0.705	0.727	1.224
Hybrid ViT	ViT-ConstantQ	0.674	0.705	1.240
Hybrid ViT	ViT-MelSpectrogram	0.664	0.697	1.127
End-to-End	RawNet2	0.659	0.663	0.731
Classical ML	SVM Baseline (csun22)	0.232	0.149	0.353

A critical finding of this study is the significant disparity in detection rates across different synthesis engines, highlighting a severe "generalization gap." The ViT models demonstrated high efficacy against open-source, zero-shot, and autoregressive models. Specifically, the ViT-MelSpectrogram model achieved detection rates of 84.6% against XTTS and 81.6% against Tortoise. However, as shown in Table 2, the detectors struggled significantly against proprietary commercial models and retrieval-based systems. Against ElevenLabs and RVC, the accuracy of the ViT models dropped to approximately 51–53%, which is statistically equivalent to random guessing. This contrast quantitatively demonstrates the generalization gap.

Table 2
Accuracy breakdown by synthesis engine using the "Different Text" (DT) datasett

Detection Model	ElevenLabs	RVC	Tortoise	XTTS
ViT-MelSpectrogram	0.516	0.516	0.816	0.846
ViT-MFCC	0.533	0.533	0.756	0.780
RawNet2 (End-to-End)	0.490	0.490	0.379	0.563

However, the detectors struggled significantly against proprietary commercial models and retrieval-based systems. Against ElevenLabs and RVC (Retrieval-based Voice Conversion) samples, the accuracy of the best-performing ViT models dropped to approximately 51–53%, which is statistically equivalent to random guessing. For instance, the ViT-MFCC model, despite its high general accuracy, achieved only 53% accuracy against ElevenLabs and RVC.

This disparity suggests that current detectors heavily overfit to the specific acoustic artifacts of open-source architectures (like the autoregressive decoding of Tortoise). In contrast, the artifacts produced by the likely diffusion-based or highly optimized vocoders of commercial "black box" engines remain imperceptible to models not explicitly trained on them.

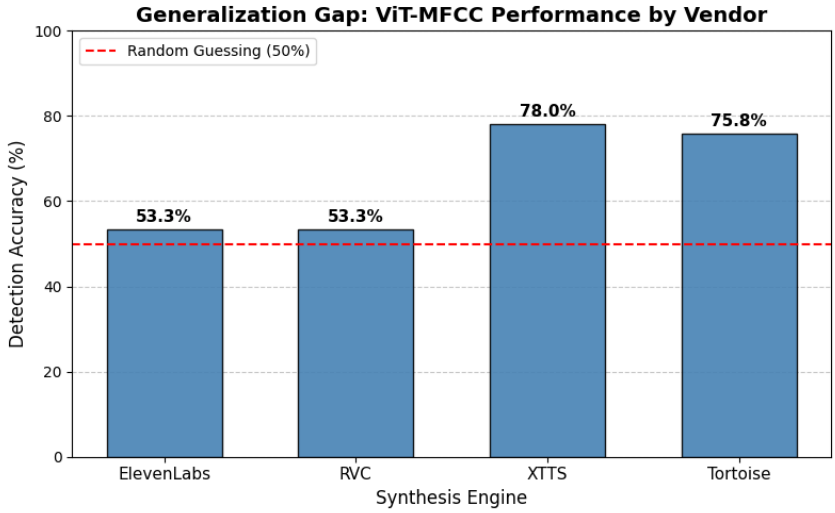

Figure 5: Detection accuracy of the ViT-MFCC model across distinct synthesis engines

The chart illustrates the significant generalization gap: while the model achieves high accuracy on open-source autoregressive engines (XTTS, Tortoise), performance degrades to near-random levels (~53%) against proprietary (ElevenLabs) and retrieval-based (RVC) systems.

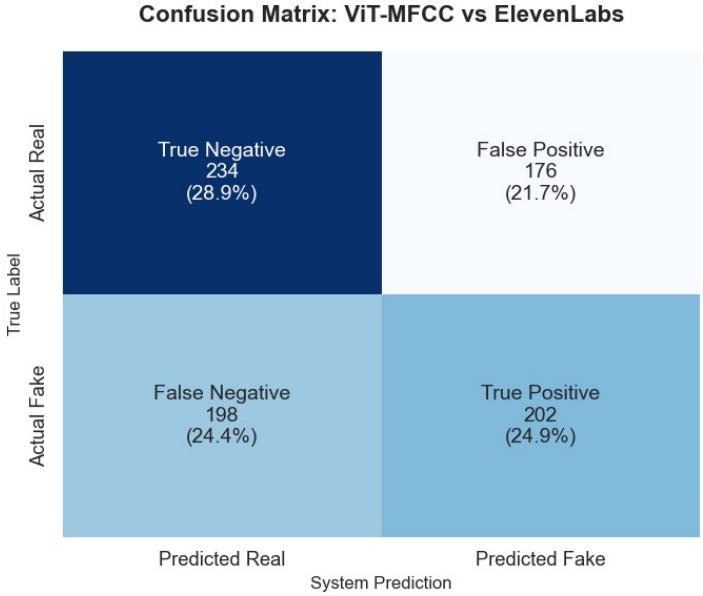

Figure 6: Confusion matrix of the ViT-MFCC model evaluated against ElevenLabs samples

The heatmap highlights a critical vulnerability: 49.5% of synthetic samples were misclassified as real speech (Missed Fakes). This high False Negative rate confirms the model's inability to generalize to unseen proprietary synthesis artifacts compared to its performance on open-source models.

4.1. Semantic Bias Assessment

To determine whether the models were learning acoustic "fakeness" or merely memorizing semantic patterns (e.g., specific phrases used in training), we compared performance on the "Different Text" (DT) and "Same Text" (ST) datasets.

- ViT Architectures: The ViT-MFCC model showed a noticeable performance drop from 0.705 (DT) to 0.665 (ST). Similarly, the ViT-MelSpectrogram model dropped from 0.664 to 0.634. This fluctuation suggests a degree of overfitting to semantic or phonetic content present in the varied text samples.
- RawNet2 Architecture: The End-to-End RawNet2 model exhibited greater stability, with accuracy shifting only slightly from 0.659 (DT) to 0.644 (ST).

This stability indicates that the End-to-End approach, which processes raw waveforms, is less susceptible to semantic bias and focuses more reliably on low-level signal artifacts, making it a potentially more robust candidate for generalized detection despite its slightly lower peak accuracy.

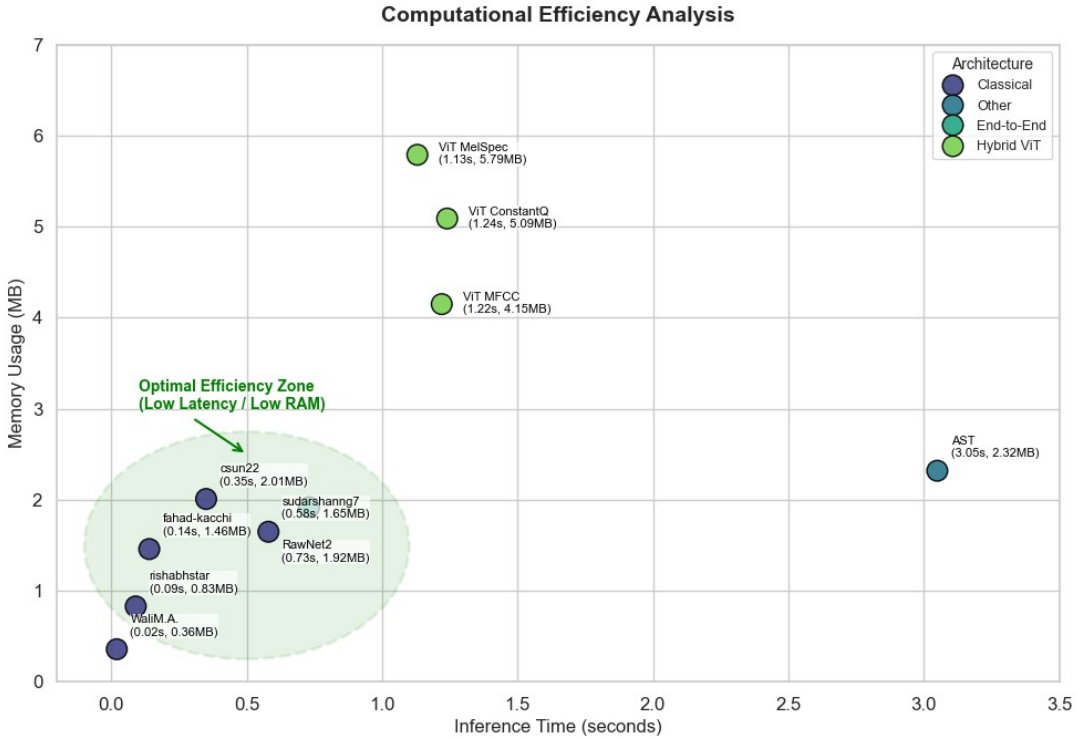

Figure 7: Computational efficiency trade-off analysis

However, detection performance is only one aspect of practical utility. To assess the viability of these models for real-time applications and IoT environments, feasibility for edge deployment was evaluated by measuring average inference time (per sample) and RAM consumption.

- Efficiency: The RawNet2 model proved to be highly efficient, requiring only 0.73 seconds for inference and consuming 1.92 MB of RAM.

- Trade-off: The ViT-MFCC model, while most accurate, required 1.22 seconds and 4.15 MB of RAM. The ViT-MelSpectrogram variant was the most resource-intensive, consuming 5.78 MB of RAM.
- Baseline: While the Classical approaches were the fastest (e.g., 0.02s to 0.35s), their inability to detect fraud renders this speed irrelevant.

These results position RawNet2 as the optimal candidate for resource-constrained environments (IoT, mobile), while ViT-MFCC offers a balanced solution for server-side deployment where detection accuracy is paramount.

The scatter plot demonstrates that RawNet2 (green) offers the optimal balance for resource-constrained environments (IoT/Edge), requiring significantly less memory (~1.92 MB) and processing time (~0.73s) compared to the more accurate but resource-intensive ViT architectures (blue).

5. Conclusions

The study concludes that despite significant advancements in deep learning architectures, a reliable and precise detection mechanism for modern neural voice cloning does not currently exist. The comparative analysis revealed that none of the evaluated systems - Classical, Hybrid ViT, or End-to-End RawNet2 - achieved the accuracy required for deployment in critical security environments. While the Hybrid ViT-MFCC model demonstrated superior performance on open-source generators, its detection capability against state-of-the-art commercial engines (ElevenLabs) and retrieval-based models (RVC) collapsed to ~51–53%, rendering it statistically equivalent to random guessing. This exposes a fundamental generalization gap: current detectors overfit to known synthesis artifacts and fail to identify the sophisticated traces of modern diffusion-based "black box" systems. Consequently, even computationally efficient architectures like RawNet2 remain impractical for biometric security without foundational improvements in accuracy. Future research may therefore prioritize adversarial robustness and data diversity, specifically integrating samples from proprietary commercial engines to force models to learn universal anomaly patterns rather than algorithm-specific artifacts.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] A. Handa. AI enabled voice cloning. FinTalk, Bank of Baroda 1–2, 2020.

[2] A. Gomez-Alanis, J. A. Gonzalez-Lopez, A. M. Peinado. Adversarial transformation of spoofing attacks for voice biometrics. arXiv preprint, 2021, arXiv:2201.01226.

[3] O. Harasymchuk, et al. Modern methods of ensuring information protection in cybersecurity systems using artificial intelligence and blockchain technology. Technology Center PC, 2025, doi:10.15587/978-617-8360-12-2

[4] CXToday: Voice cloning tech is breaking customer authentication systems, 2024. https://www.cxtoday.com/contact-centre/voice-cloning-tech-is-breaking-customer-authentication-systems/

[5] X. Tan, et al. A survey on neural speech synthesis. arXiv preprint, 2021, arXiv:2106.15561

[6] V. Poberezhnyk, V. Balatska, I. Opirskyy. Development of the Learning Management System Concept based on Blockchain Technology. CEUR Workshop Proceedings, 3550 (2023) 143–156.

[7] P. Petriv, I. Opirskyy, N. Mazur. Modern technologies of decentralized databases, authentication, and authorization methods. Cybersecurity providing in information and telecommunication systems, 3826 (2024) 60–71.
[8] J. Betker. Better speech synthesis through scaling. arXiv preprint arXiv:2305.07243 (2023)
[9] V. Brydinskyi, et al. Enhancing Automatic Speech Recognition With Personalized Models: Improving Accuracy Through Individualized Fine-Tuning, IEEE Access, 12, 2024, 116649–116656, doi:10.1109/ACCESS.2024.3443811
[10] ElevenLabs: AI voice generation platform (2025). https://www.elevenlabs.io
[11] V. Dudykevych, et al., Detecting Deepfake Modifications of Biometric Images using Neural Networks, in: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, CPITS, vol. 3654 (2024) 391–397.
[12] O. Iosifova, et al., Techniques Comparison for Natural Language Processing, in: 2nd International Workshop on Modern Machine Learning Technologies and Data Science, vol. 2631, no. I (2020) 57–67.
[13] I. Iosifov, O. Iosifova, V. Sokolov, Sentence Segmentation from Unformatted Text using Language Modeling and Sequence Labeling Approaches, in: VII International Scientific and Practical Conference Problems of Infocommunications. Science and Technology (2020) 335–337. doi:10.1109/PICST51311.2020.9468084
[14] I. Iosifov, et al., Transferability Evaluation of Speech Emotion Recognition Between Different Languages, Advances in Computer Science for Engineering and Education 134 (2022) 413–426. doi:10.1007/978-3-031-04812-8_35
[15] I. Iosifov, et al., Natural Language Technology to Ensure the Safety of Speech Information, in: Workshop on Cybersecurity Providing in In-formation and Telecommunication Systems, vol. 3187, no. 1 (2022) 216–226.
[16] O. Romanovskyi, et al., Accuracy improvement of spoken language identification system for close-related languages, Advances in Computer Science for Engineering and Education VII, vol. 242 (2025) 35–52. doi:10.1007/978-3-031-84228-3_4
[17] Coqui AI: XTTS. Open multilingual zero-shot text-to-speech model (2025). https://github.com/coqui-ai/xtts
[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
[19] Ali, S., Tanweer, S., Khalid, S.S., Rao, N.: Overview of Mel-Frequency Cepstral Coefficients (MFCC). In: Proceedings of the 2nd Int. Conf. on Digital, Smart and Sustainable Development (ICIDSSD), EAI, 2020.
[20] J. C. Brown. Calculation of a constant Q spectral transform. Journal of the Acoustical Society of America 89(1), 1991, 425–434.
[21] J. W. Jung, H. S. Heo, H. J. Shim, H. J. Yu. RawNet2: End-to-end neural speaker verification with raw waveform input. arXiv preprint, 2021, arXiv:2011.01108
[22] I. Zaiets, et al. Integrated System for Speaker Diarization and Intruder Detection using Speaker Embeddings, in: CEUR Workshop Proceedings, 3654 (2024) 228–238.
[23] F. Pedregosa, et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12 (2011) 2825–2830.
[24] MattyB95: Jabberjay. Deep Fake Audio Detection Framework, 2025. https://github.com/MattyB95/Jabberjay
[25] Sun, C.: Synthetic Voice Detection. Vocoder Artifacts (csun22), 2025. https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts