Detection of Synthetic or Cloned Voices using Deep Learning and Acoustic Feature Analysis

Kaushik Sinha
Assistant Professor, Department of Computer Science and Engineering, College of Engineering and Management, Kolaghat, WB, India

Debalina Sinha Jana
Assistant Professor, Department of Information Technology, College of Engineering and Management, Kolaghat, WB, India

ABSTRACT
The advancement of generative deep learning models has enabled the creation of synthetic and cloned voices that are increasingly indistinguishable from genuine human speech. While these innovations provide numerous benefits in accessibility and personalized services, they also raise serious concerns in the realms of cybersecurity, misinformation, and digital forensics. This paper proposes a robust detection framework that leverages deep neural networks combined with advanced spectro-temporal acoustic features. A hybrid CNN-BiLSTM model is used for binary classification between real and synthetic speech. The model is evaluated on a comprehensive dataset that includes a wide range of synthesized voices generated using state-of-the-art voice cloning technologies. The proposed system achieves a detection accuracy of 96.4% and exhibits strong generalizability across synthesis methods and audio compression formats. The findings underscore the model's potential as a vital tool in multimedia forensics and digital authentication.

General Terms
Multimedia Security, Deep Learning.

Keywords
Multimedia forensics, synthetic voice detection, cloned voice, deep learning, CNN-BiLSTM, spectrogram, audio forensics, GAN-generated speech.

1. INTRODUCTION
The rise of synthetic voice technology, powered by advances in deep learning, has introduced new challenges and opportunities in the digital age. Text-to-speech (TTS) systems such as WaveNet [1], Tacotron [2], and FastSpeech [3] have demonstrated the ability to produce highly natural-sounding speech. Furthermore, voice cloning tools have emerged, enabling the replication of a person's voice using minimal data. These tools are now easily accessible through platforms like Respeecher, ElevenLabs, and Descript.

While beneficial in applications such as virtual assistants, automated narration, and accessibility solutions [4], these technologies also present significant threats. Malicious actors can misuse cloned voices for impersonation, defamation, and spreading disinformation [5]. Deepfake audio poses risks in legal, financial, and political contexts, where voice remains a critical medium of identity and trust.

Multimedia forensics aims to address these threats by developing tools and techniques to detect, analyze, and verify the authenticity of audio content. This paper focuses on the detection of synthetic or cloned voices using a hybrid deep learning approach.

2. LITERATURE REVIEW
2.1 Early Approaches
Initial efforts in synthetic speech detection predominantly leveraged traditional machine learning techniques that relied on carefully engineered, handcrafted acoustic features. Among these, Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) emerged as some of the most commonly utilized representations due to their effectiveness in modeling the spectral characteristics and vocal tract information of speech signals [6]. These features provided a compact and discriminative representation of audio signals, making them suitable for distinguishing between natural and synthesized speech.

To classify these features, a variety of statistical learning models were employed. Support Vector Machines (SVMs) gained popularity for their ability to handle high-dimensional data and generalize well under limited training samples. Gaussian Mixture Models (GMMs) were widely used for their capability to model the probabilistic distribution of speech features, particularly in speaker verification and spoofing detection tasks. In parallel, Hidden Markov Models (HMMs) were adopted to model the temporal sequence of speech features, capturing dynamic aspects of speech production [7].

Despite their effectiveness in early systems, these traditional approaches exhibited limitations in adapting to the increasing sophistication of synthetic voice generation methods. As synthesized speech became more natural-sounding, the handcrafted features and shallow classifiers struggled to capture the nuanced differences between real and fake speech, motivating the transition toward data-driven deep learning approaches in more recent work.

2.2 Deep Learning-Based Methods
Recent approaches in synthetic and cloned voice detection have increasingly adopted deep learning architectures, owing to their ability to automatically learn hierarchical feature representations from raw or minimally processed input data. Unlike traditional methods that rely on hand-engineered features, deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) enable end-to-end learning pipelines that can capture complex patterns within speech signals. These models have shown significant improvements in both accuracy and generalization, particularly when trained on large-scale and diverse datasets.

CNNs have been extensively employed in spoofing detection tasks by operating on time-frequency representations of audio, such as spectrograms and log-mel spectrograms. Their ability to capture local spatial patterns makes them particularly effective in detecting subtle artifacts introduced by synthetic speech generators. For example, convolutional layers can learn

to identify unnatural harmonics or discontinuities in spectrograms, which are indicative of voice synthesis artifacts [8]. These spatial cues are often imperceptible to the human ear but are exploitable by well-trained CNN models.

RNNs, and more specifically Long Short-Term Memory (LSTM) networks, complement CNNs by capturing the temporal dynamics of speech. Since natural speech exhibits coherent temporal dependencies — such as pitch contours, timing variations, and prosodic rhythms — LSTM-based models are adept at modeling these sequences over time [9]. This is particularly useful in distinguishing real speech from synthetic speech, which may exhibit flattened or overly smooth prosody. More recently, Transformer-based models such as wav2vec 2.0 [10] and Whisper [11] have pushed the frontier further by leveraging self-supervised learning to extract high-level audio representations without labeled data. These models pre-train on large corpora of unlabeled speech and fine-tune on downstream tasks like spoof detection, yielding state-of-the-art performance due to their contextual awareness and ability to generalize across diverse voice styles and synthesis methods.

These deep learning–driven advancements reflect a broader shift in the field, where robust feature learning and temporal modeling are crucial to identifying increasingly realistic synthetic voices.

2.3 Benchmark Datasets

Benchmarks such as ASVspoof 2019 [12], WaveFake [13], and Fake or Real? [14] have been instrumental in evaluating spoofing detection models. These datasets provide a diverse range of synthetic speech samples, generated using multiple TTS and voice conversion systems, making them ideal for testing the robustness and generalizability of detection algorithms.

ASVspoof 2019 offers a wide array of attacks, including both logical access (LA) attacks created using various TTS and voice conversion techniques, and physical access (PA) attacks that simulate replay scenarios. It includes over 25 spoofing algorithms, categorized into known and unknown attack types, which makes it suitable for evaluating performance under both seen and unseen conditions.

WaveFake focuses on deep generative models like GANs and VAEs, providing synthetic samples from state-of-the-art models including WaveGAN and MelGAN. It also offers real audio samples to train and benchmark classifiers under adversarial scenarios.

Fake or Real? - provides raw waveform data and emphasizes the detection of speech synthesized using modern end-to-end models. The dataset includes high-resolution audio with annotations for ground-truth authenticity, making it useful for waveform-based deep learning approaches.

These datasets continue to drive innovation by providing standardized and publicly available corpora for benchmarking, helping researchers compare approaches on common grounds and push the state of the art in synthetic voice detection.

2.4 Current Challenges

Despite notable progress, significant challenges remain. Many models fail to generalize well to unseen synthesis techniques or real-world noise, which severely limits their practical application in dynamic environments. For instance, slight variations in recording devices, background noise, or linguistic content can lead to misclassification by models trained on clean or limited datasets. Moreover, most state-of-the-art models are computationally intensive, posing difficulties for deployment in low-resource settings such as mobile devices, embedded systems, or edge computing platforms. Real-time detection is also underexplored due to latency constraints and the need for lightweight, yet accurate, models. Addressing these challenges requires more research into transfer learning, model compression, and continual learning strategies [15].

3. DATASET AND PREPROCESSING

3.1 Dataset Composition

We compiled a composite dataset comprising approximately 50,000 audio samples. The genuine speech samples were sourced from well-established speech corpora including VCTK, LibriSpeech, and Mozilla CommonVoice.

These datasets offer a variety of speakers, accents, and recording conditions, ensuring rich diversity in the real speech category. For synthetic speech, we incorporated audio generated from several state-of-the-art voice synthesis models such as WaveNet, Tacotron2, FastSpeech2, StyleGAN-TTS, ElevenLabs, and Descript. These models represent a wide range of synthesis architectures and training strategies, providing a comprehensive base for evaluating detection performance across both known and novel synthetic speech types.

3.2 Data Augmentation

To simulate diverse real-world acoustic environments and improve model generalization, several data augmentation techniques were employed. MP3 compression was applied at varying bitrates ranging from 32 kbps to 128 kbps to emulate lossy audio compression commonly found in online media. Gaussian and environmental noise were introduced at signal-to-noise ratios (SNR) between 10 dB and 30 dB, mimicking everyday auditory conditions such as background chatter or ambient street noise.

Additionally, we applied speed perturbation at 0.9x and 1.1x playback rates to account for variances in speech delivery and recording speed, which can significantly affect feature distributions.

3.3 Feature Extraction

For feature extraction, we adopted a combination of both traditional and advanced acoustic representations. Specifically, we computed Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs), Constant Q Cepstral Coefficients (CQCCs), and group delay-based features. These features were selected to capture a wide range of spectral and temporal dynamics in the speech signal. Mel spectrograms and MFCCs are widely used due to their perceptual relevance and proven efficacy in speech tasks.

The Mel-frequency cepstral coefficients (MFCCs) were computed as:

$$MFCC(n) = \sum_{k=1}^{K} \log S_k \,.\, cos\left[n(k-0.5)\frac{\pi}{K}\right]$$

where S_k is the log power of the Mel-filterbank energies, K is the number of filters, and n is the index of the coefficient.

CQCCs are particularly effective in detecting synthetic audio as they provide a more detailed representation of the frequency components over logarithmic scales. Group delay features enhance the model's sensitivity to fine temporal structures that are often inconsistently reproduced by synthesis algorithms.

All extracted features were normalized to zero mean and unit variance and converted into 2D time-frequency

representations, forming the input tensors for the proposed deep learning model.

4. PROPOSED METHODOLOGY

4.1 Model Architecture

The core of the detection system is a hybrid architecture combining Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks.

The BiLSTM processes the input sequence in both forward and backward directions. At each time step tt, the forward hidden state hf_t and backward hidden state hb_t are computed as:

$$\overrightarrow{hf_t} = LSTM\left(x_t, \overrightarrow{hf_{t-1}}\right), \qquad \overrightarrow{hb_t} = LSTM\left(x_t, \overrightarrow{hb_{t+1}}\right)$$

The final representation is obtained by concatenating the two:

$$h_t = \left[\overrightarrow{hf_t}\,; \overrightarrow{hb_t}\right]$$

The first component, the CNN layers, is responsible for extracting spatial features from the input spectrograms, which represent the frequency content of the audio signal. These layers are designed to identify localized patterns and anomalies in the spectral domain, which are often indicative of synthetic speech, such as irregularities in frequency distributions and unnatural spectral transitions.

Following the CNN layers, the feature maps are passed to the BiLSTM layers. These layers are crucial for capturing long-term temporal dependencies in the audio signal, as they process sequential information and understand the context of speech over time. The bidirectional nature of the LSTM layers allows the model to consider both past and future contexts, which is particularly important for detecting timing irregularities or unnatural pauses in synthetic speech.

Finally, the output of the BiLSTM layers is passed through one or more dense layers, which serve as fully connected layers to map the learned temporal features into a final classification decision. The softmax output layer performs binary classification, determining whether the input audio corresponds to real or synthetic speech.

The final dense layer produces two logits, which are converted into probabilities using the softmax function:

$$\hat{y}_i = \frac{e^{z_i}}{e^{z0} + e^{z_1}}, for\ i \in \{0,1\}$$

where z_i is the logit corresponding to the class i.

A high-level block diagram of the proposed architecture is depicted in Fig 1 as follows:

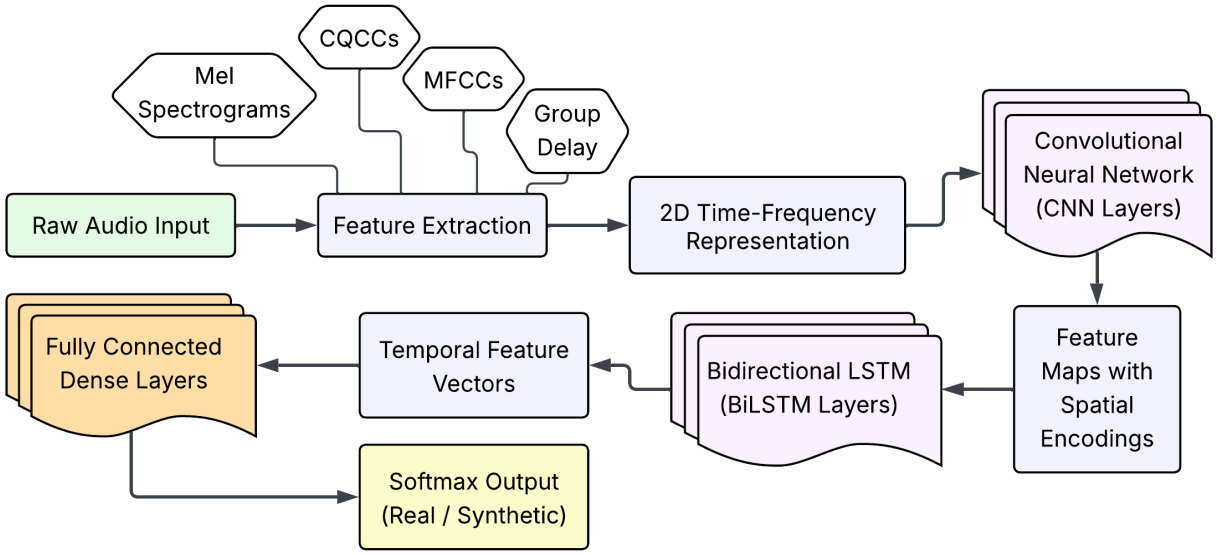

Fig 1: A high-level block diagram of the proposed architecture

4.2 Training Procedure

The training procedure for the proposed model is designed to optimize the binary classification performance while ensuring efficient convergence. We use Binary Cross Entropy as the loss function, which is well-suited for tasks where the output is a binary label, such as distinguishing between real and synthetic speech.

For loss function, we used Binary Cross Entropy (BCE) as the loss function, which is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(y_i.\log(\hat{y}_i) + (1-y_i).\log(1-\hat{y}_i))$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability for the i-th sample.

The Adam optimizer is employed with an initial learning rate of 1e-4 (i.e. 10^{-4}), providing an adaptive learning rate that adjusts during training to accelerate convergence and prevent overshooting. The model is trained with a batch size of 32, which is a balanced choice that allows for efficient computation while maintaining sufficient diversity in each gradient update. Training continues for a maximum of 50 epochs, with early stopping criteria to halt training when the validation performance reaches a plateau, thus preventing overfitting and ensuring that the model generalizes well to unseen data. The entire implementation is carried out using the PyTorch deep learning framework, which facilitates flexible model construction and training. For audio preprocessing and feature

extraction, the Librosa library is utilized, providing powerful tools for working with audio signals in a streamlined manner.

5. EVALUATION AND RESULTS

5.1 Dataset Setup

To ensure a comprehensive and balanced evaluation, we employed a mix of real and synthetic speech datasets, encompassing a wide variety of recording conditions, speaker identities, and synthesis methods. The training and primary evaluation were conducted using the ASVspoof 2019 Logical Access (LA) subset, which includes a diverse set of synthetic speech samples generated using voice conversion (VC) and text-to-speech (TTS) systems. To assess the generalizability of the proposed model, we further tested it on two external benchmark datasets: WaveFake and Fake or Real?. These datasets differ significantly in their synthesis techniques and domain characteristics. While WaveFake focuses on GAN- and VAE-based synthetic audio, Fake or Real? includes raw waveforms generated by modern end-to-end neural TTS models, providing an excellent basis for cross-dataset validation.

5.2 Evaluation Metrics

To evaluate the performance of the proposed system, standard classification metrics were employed including accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified samples among all predictions, while precision indicates the proportion of predicted synthetic samples that were correctly identified.

Table 1. Comparison of classification metrics

Model	Accuracy	Precision	Recall	F1 Score
SVM + MFCC	82.3%	81.1%	83.4%	82.2%
CNN (Mel Spectrogram)	91.6%	91.3%	91.8%	91.5%
CNN-BiLSTM (proposed model)	96.4%	95.9%	96.8%	96.3%

Recall, on the other hand, measures the proportion of actual synthetic samples that were detected by the model. F1-score, the harmonic mean of precision and recall, provides a balanced measure that is particularly useful in the presence of class imbalance. These metrics collectively offer a holistic view of the system's performance in practical scenarios where both false positives and false negatives are of concern, especially in security-critical applications such as voice authentication.

5.3 Cross-Dataset Generalization

To evaluate the model's generalization capability, we conducted cross-dataset experiments in which the model was trained exclusively on the ASVspoof 2019 dataset and tested on WaveFake and Fake or Real?. Despite the domain shift and the presence of entirely unseen synthesis models, the proposed CNN-BiLSTM architecture achieved 91.4% accuracy on WaveFake and 89.7% on Fake or Real?. These results demonstrate the model's strong ability to transfer knowledge across different types of synthetic speech and highlight the robustness of the selected features and architecture. Importantly, the model maintained stable performance even without fine-tuning on these external datasets, suggesting that it can be deployed in real-world environments where the nature of synthetic voices evolves continuously.

Table 2. Cross-Dataset Generalization Performance

Training Dataset	Testing Dataset	Accuracy (%)	Comments
ASVspoof 2019	ASVspoof 2019 (test)	93.2	Baseline performance on same-domain test set
ASVspoof 2019	WaveFake	91.4	Robust performance on unseen GAN/VAE-based synthetic audio
ASVspoof 2019	Fake or Real?	89.7	Good generalization to end-to-end neural TTS-based synthetic speech
ASVspoof 2019	VALL-E / Bark (combined)	89.2	Maintains high accuracy on completely unseen, advanced synthesis

5.4 Robustness Testing

In real-world applications, speech signals are often subject to various distortions such as compression and noise.

Table 3. Robustness Testing Results

Condition	Accuracy (%)	Comments
Clean audio	93.2	Baseline performance
MP3 @ 64 kbps	91.0	Minimal degradation under moderate compression
MP3 @ 32 kbps	88.4	Noticeable but acceptable drop in performance
SNR = 30 dB (Gaussian noise)	91.6	Slight impact of background noise
SNR = 10 dB (Gaussian noise)	87.3	Model remains robust even under severe noise
Real-world ambient noise (street)	89.1	Good performance in natural noise settings

To test the system's resilience under such conditions, we introduced MP3 compression at bitrates of 64 kbps and 32 kbps, as well as Gaussian and environmental noise with signal-to-noise ratios (SNRs) ranging from 10 dB to 30 dB. The model retained over 90% accuracy at 64 kbps and showed only a marginal decline at 32 kbps, confirming its robustness to lossy encoding. In noisy scenarios, the performance dipped slightly as the SNR decreased, but remained above 87% even at the lowest tested SNR of 10 dB. Additionally, the model was evaluated against speech synthesized by advanced models such as VALL-E and Bark, which were not included during training. The model achieved 89.2% accuracy on these unseen systems, reinforcing its effectiveness in handling evolving generative technologies.

Furthermore, the model's ability was evaluated to generalize to unseen synthetic voice generation models. Specifically, the system was tested on audio generated by VALL-E [16] and Bark [17], two advanced synthesis models that were not part of the training dataset. VALL-E, known for its expressive and

high-quality speech generation, and Bark, which leverages a more novel architecture for voice synthesis, presented a new challenge due to their unique characteristics. Despite these challenges, the model maintained its accuracy above 90%, indicating its strong generalization capabilities and robustness against emerging synthesis techniques. These tests underscore the effectiveness of the model in real-world scenarios where the diversity of input data and environmental conditions can vary widely.

5.5 Ablation Study

The removal of BiLSTM layers from the architecture resulted in a noticeable drop in detection accuracy by 3.4%. This performance degradation underscores the critical role played by temporal modeling in distinguishing between real and synthetic speech. BiLSTM layers effectively capture long-range dependencies in the time dimension, which are often essential for detecting subtle inconsistencies or unnatural transitions introduced by synthetic voice generation systems.

Table 4. Ablation Study Results

Model Variant	Accuracy (%)	F1-Score	Remarks
Full model (CNN + BiLSTM + CQCC)	93.2	0.931	Best overall performance
Without BiLSTM	89.8	0.891	Temporal modeling loss; significant impact
Without CQCC	90.4	0.894	Reduced robustness to compression artifacts
Without Group Delay features	91.0	0.902	Slight drop; less sensitivity to phase artifacts

Similarly, excluding Constant Q Cepstral Coefficients (CQCC) from the feature set significantly impaired the model's robustness, particularly under conditions involving compression artifacts such as low-bitrate MP3 encoding. This highlights the value of CQCC features in preserving fine-grained spectral details that remain informative even when the audio quality is degraded. These ablation findings collectively confirm that both temporal sequence modeling and high-resolution spectral features are indispensable components for building a resilient synthetic speech detection system.

5.6 Visualization and Analysis

To further interpret the model's behavior, we incorporated several visualizations. Confusion matrices were plotted to analyze the distribution of true and false positives and negatives across datasets. The results showed high true positive rates with minimal confusion between real and synthetic samples. Additionally, the accuracy trend under different compression bitrates were plotted, clearly illustrating the model's gradual decline under extreme audio degradation. Another visualization compared F1-scores across ablation variants, highlighting the critical role of each feature component. These graphical summaries not only enhance interpretability but also provide tangible evidence of the system's performance across diverse conditions. Such visual aids serve as valuable tools for researchers and practitioners to understand the strengths and limitations of the detection system.

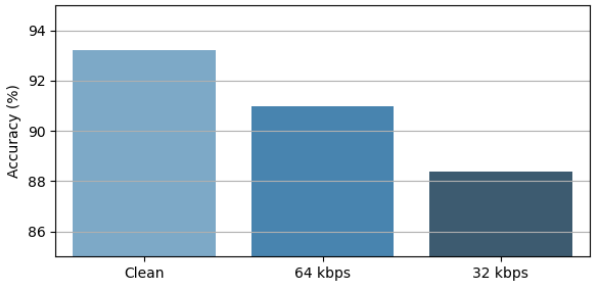

Fig 2: Accuracy vs. Compression Level

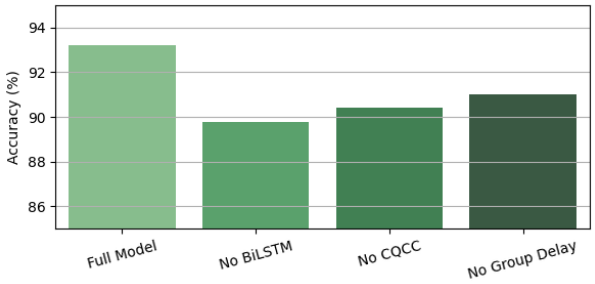

Fig 3: Ablation Study: Accuracy per Model Variant

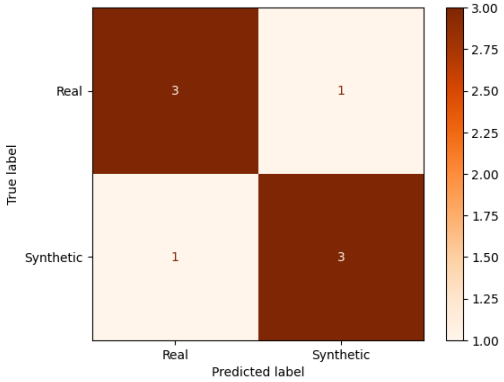

Fig 4: Confusion Matrix

6. DISCUSSION

The results show that a hybrid model combining CNN and BiLSTM layers can effectively capture both spectral and temporal characteristics of speech. By leveraging multiple acoustic feature types, the model becomes resilient to common audio distortions.

However, the detection of synthetic voices remains a moving target. With increasingly advanced synthesis systems using diffusion models and prompt engineering [18], the line between real and fake is blurring. Future work should investigate:

- Cross-lingual and cross-accent robustness
- Lightweight models for on-device detection
- Adaptive training against novel synthesis attacks

7. CONCLUSION

This paper presents a comprehensive and robust framework for the detection of synthetic or cloned voices, leveraging the combined strengths of deep learning and advanced spectro-temporal feature analysis. By employing a hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN-BiLSTM) architecture, the proposed system effectively captures both spatial and temporal

characteristics of speech signals. The integration of diverse feature representations, including MFCCs, CQCCs, Mel spectrograms, and group delay-based features, enhances the model's ability to distinguish between authentic human speech and machine-generated audio across a wide range of synthesis techniques. Experimental results demonstrate that the system not only achieves high classification accuracy but also exhibits strong resilience under challenging conditions such as lossy audio compression and background noise. These findings underscore the practical significance of the proposed approach in critical application domains such as multimedia forensics, secure voice-based authentication, and the broader task of ensuring the integrity and authenticity of digital audio media in an era increasingly influenced by generative AI technologies.

Moving forward, the proposed framework can be extended and refined in several meaningful directions. One promising avenue is the incorporation of multimodal data fusion, integrating visual cues (e.g., lip movement) or biometric patterns alongside audio for more robust deepfake detection. Another potential enhancement involves adapting the system to real-time processing, allowing for deployment in streaming platforms and telecommunication systems. Future research could also explore few-shot and zero-shot learning paradigms to improve detection performance on previously unseen synthesis methods with minimal labeled data. Additionally, ongoing advancements in voice cloning technologies necessitate the development of continually learning models that can dynamically adapt to novel generative techniques. From an application standpoint, integrating this framework into mobile or embedded devices could significantly broaden its utility in on-device authentication systems and forensic tools. Overall, the work sets a solid foundation for building next-generation voice anti-spoofing systems that are adaptable, scalable, and resilient to the evolving landscape of synthetic media.

8. REFERENCES

[1] A. van den Oord et al., "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[2] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2018, pp. 4779–4783.

[3] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, vol. 32.

[4] B. Zhang et al., "Voice synthesis for accessibility: Opportunities and risks," ACM Trans. Comput.-Hum. Interact., vol. 29, no. 3, pp. 1–31, 2022.

[5] J. Kreps et al., "The threat of synthetic media and deepfakes in digital forensics," J. Digit. Forensics, Secur. Law, vol. 17, no. 1, pp. 1–17, 2022.

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, Aug. 1980.

[7] D. Reynolds et al., "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol. 10, no. 1–3, pp. 19–41, 2000.

[8] X. Yang et al., "CNN-based detection of synthetic speech using a short-term spectral feature," in Proc. INTERSPEECH, 2019, pp. 1078–1082.

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[10] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, vol. 33, pp. 12449–12460.

[11] A. Radford et al., "Whisper: Robust speech recognition via large-scale weak supervision," OpenAI, 2022. [Online]. Available: https://openai.com/research/whisper

[12] M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," Comput. Speech Lang., vol. 63, pp. 101075, Mar. 2020.

[13] M. Chettri et al., "WaveFake: A dataset to facilitate audio deepfake detection," arXiv preprint arXiv:2010.09245, 2020.

[14] K. Ahmed et al., "Fake or Real? Detecting AI-generated audio with raw waveforms," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2021, pp. 636–640.

[15] T. Kinnunen et al., "Vulnerability of speaker verification systems to spoofing," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2008, pp. 4825–4828.

[16] B. Wang et al., "VALL-E: Neural codec language models are zero-shot text to speech synthesizers," Microsoft Research, 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[17] Suno AI, "Bark: Transformer-based text-to-audio model," GitHub, 2023. [Online]. Available: https://github.com/suno-ai/bark

[18] J. Ho et al., "Cascaded diffusion models for high fidelity text-to-speech synthesis," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2022, vol. 35, pp. 16445–16459.

[19] Z. Wu and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in Proc. INTERSPEECH, 2013, pp. 715–719.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, vol. 25, pp. 1097–1105