Mel-Spectrogram Representations for CNN Based Adversarial Defense for Audio Deepfake Detection

Iftikhar Ahmed
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
iftikhar.ahmed@northsouth.edu

Mueid Islam Arian
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mueid.arian@northsouth.edu

Nusrat Jahan Lisa
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
nusrat.lisa@northsouth.edu

***Abstract*—Audio deepfakes generated by neural Voice Encoders (vocoders) pose escalating threats to authentication systems, misinformation detection, and media integrity. While Linear Frequency Cepstral Coefficients (LFCC) have emerged as the standard feature representation for audio deepfake detection based on Gaussian Mixture Model (GMM) classifiers, their suitability for modern deep convolutional neural networks remains unexplored. This paper presents systematic comparisons of LFCC versus Mel-Spectrogram representations specifically for CNN-based detection, investigating both detection accuracy and adversarial robustness. We evaluated the Xception architecture with ImageNet pretraining on the WaveFake dataset, conducting leave-one-out cross-validation across seven neural vocoders. Findings demonstrate that Mel-Spectrograms significantly outperform LFCC for CNN-based detection, achieving 0.034 Equal Error Rate (EER) compared to 0.053 for LFCC a 36% improvement that establishes new state-of-the-art performance, surpassing both GMM baselines and RawNet2. Furthermore, we implement adaptive adversarial training and demonstrate that CNN-based detectors can achieve robust performance under white-box attacks: Fast Gradient Sign Method (FGSM) attack EER improves from 0.70 to 0.107 (86% improvement) and Projected Gradient Descent (PGD) attack EER from 0.90 to 0.154. These findings challenge the prevailing assumption that LFCC superiority generalizes from GMMs to CNNs, revealing that feature representation optimization depends critically on classifier architecture.**

***Index Terms*—Audio Deepfake Detection, Mel-Spectrogram, Linear Frequency Cepstral Coefficients, Xception Network, Adversarial Training, Neural Vocoders, Convolutional Neural Networks, Adversarial Robustness**

I. Introduction

Neural vocoders have transformed speech synthesis, enabling applications such as virtual assistants, human–computer interaction systems, and assistive technologies for individuals with speech impairments. State-of-the-art models such as MelGAN, HiFi-GAN, WaveGlow, and Parallel WaveGAN (PWG) generate highly natural speech that is often indistinguishable from genuine recordings [1]–[4]. While these advances improve accessibility and quality, they also introduce serious security risks. High-quality audio deepfakes have been exploited for voice impersonation fraud, political misinformation, and fabricated audio evidence. A widely cited CEO voice impersonation attack resulted in a $243,000 financial loss, underscoring the urgency of robust audio deepfake detection [5].

To address these threats, the research community has developed benchmark initiatives such as the ASVspoof challenge series, which provide standardized datasets and evaluation protocols for anti-spoofing systems [6]. Early approaches relied on traditional machine learning models with handcrafted acoustic features. Using the WaveFake dataset, prior work showed that Linear Frequency Cepstral Coefficients (LFCC) outperform Mel-Frequency Cepstral Coefficients (MFCC) when paired with Gaussian Mixture Model (GMM) classifiers [7], leading to widespread adoption of LFCC in deepfake detection.

Despite this progress, it remains unclear whether the advantages of LFCC observed in statistical learning extend to modern deep Convolutional Neural Networks (CNNs), which learn hierarchical spatial patterns from dense two-dimensional inputs.

In this work, we propose an Xception-based audio deepfake detection framework [8] and conduct a controlled comparison of LFCC and Mel-Spectrogram features using identical architectures and training protocols. Our results show that Mel-Spectrogram-based models achieve a 36% reduction in Equal Error Rate compared to LFCC under CNN-based detection, demonstrate stronger robustness under FGSM and PGD attacks, and generalize better to unseen synthesis models through leave-one-out cross-validation.

The remainder of this paper is organized as follows. Section II reviews related work, Section III describes the methodology, Section IV presents results and analysis, Section V discusses the findings, and Section VI concludes.

II. Related Work

Modern audio deepfakes are generated using neural vocoders, including flow-based models (WaveGlow), GAN-based architectures (MelGAN, HiFi-GAN, Parallel WaveGAN), and diffusion models, which introduce distinct synthesis artifacts.

Frank and Schönherr demonstrated that LFCC outperforms MFCC for GMM-based detection on WaveFake, achieving an EER of 0.058 [7]. RawNet2 achieved an EER of 0.040 using end-to-end learning from raw waveforms but exhibited severe overfitting, with some held-out vocoders reaching EER = 0.985 [9]. Whether LFCC superiority extends to deep CNNs, where

979-8-3315-4864-3/26/$31.00 ©2026 IEEE

dense spatial representations and pretrained architectures may offer advantages, remains unexplored.

Recent work showed strong adversarial vulnerability, with FGSM degrading an LCNN from EER 0.022 to 0.785 [10]. Adaptive adversarial training improved robustness to EER 0.125 (86% improvement) but reduced clean performance to EER 0.088, highlighting the accuracy-robustness tradeoff. Prior studies focus on lightweight CNNs with LFCC, leaving the impact of dense spectral features and large-scale pretrained models under adversarial training as open questions.

III. Methodology

A. Dataset and Preprocessing

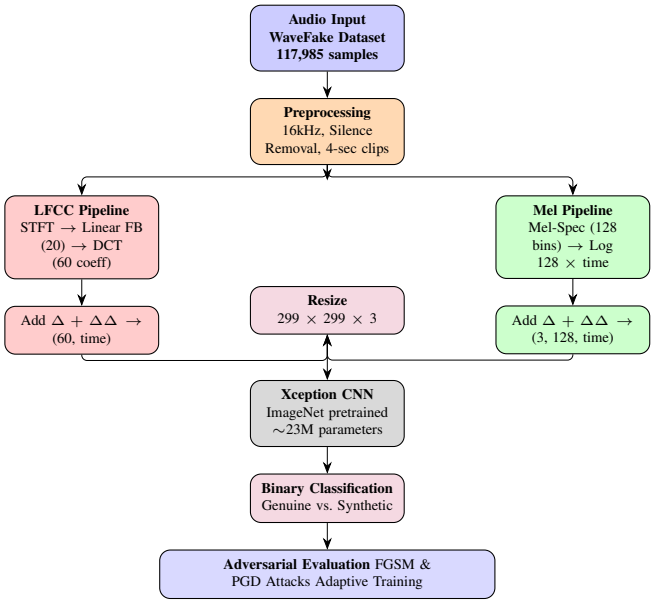

Fig. 1. End-to-end pipeline for audio deepfake detection. Audio samples are transformed into LFCC or Mel-Spectrogram representations, resized to 299×299×3, and classified using an ImageNet-pretrained Xception CNN to isolate the effect of feature representation on performance and robustness.

We employ the WaveFake dataset [7] containing 117,985 total samples (104,885 generated and 13,100 real) from six neural vocoder architectures plus full a TTS pipeline outputs, generated from LJSpeech [11]. Preprocessing standardizes all audio: resample to 16 kHz mono, remove silence (>2s gaps, 40 dB threshold), normalize duration to 4 seconds. Data splitting employs two protocols. Single-vocoder experiments partition samples into training (60%), validation (20%), and test (20%) sets to assess in-distribution performance. Leave-one-out cross-validation trains on six vocoders while testing on the seventh held-out generator, repeated across all eight synthesis methods to evaluate cross-vocoder generalization.

As illustrated in Fig. 1, the proposed pipeline preprocesses audio samples and extracts either LFCC or Mel-Spectrogram features before classification using an ImageNet-pretrained Xception CNN [12].

B. Feature Extraction

We design two parallel feature extraction pipelines based on LFCC and Mel-Spectrogram representations, illustrated in Fig. 2 and Fig. 3. Both pipelines share identical preprocessing while preserving their fundamental representational differences. Fig. 2 was generated from a single audio file following our feature extraction pipeline.

The LFCC pipeline applies Short-Time Fourier Transform ($n_{fft} = 2048$, hop=512) followed by a linear-frequency filterbank (20 filters, 0-8000 Hz) and Discrete Cosine Transform producing 20 cepstral coefficients. First and second derivatives (Δ and $\Delta\Delta$) yield $(60, \text{time})$ representations, resized to $299 \times 299 \times 3$ for CNN input.

The Mel-Spectrogram pipeline uses identical STFT parameters with mel-scaled filterbank ($n_{mels} = 128$) and log-power compression. Critically, no DCT is applied, preserving dense spectral information. Temporal derivatives form a $(3, 128, \text{time})$ tensor, resized to $299 \times 299 \times 3$.

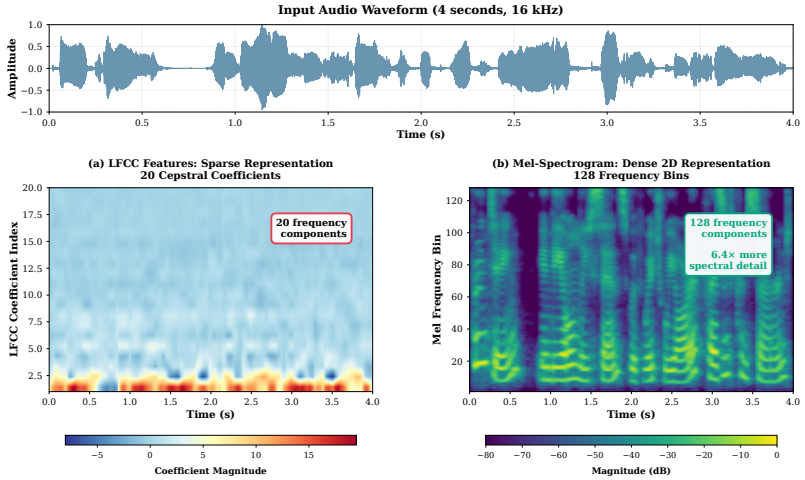

Fig. 2. LFCC and Mel-Spectrogram representations extracted from the same 4-second audio sample. LFCC produces a sparse cepstral representation, while Mel-Spectrograms preserve dense time–frequency structure using 128 frequency bins.

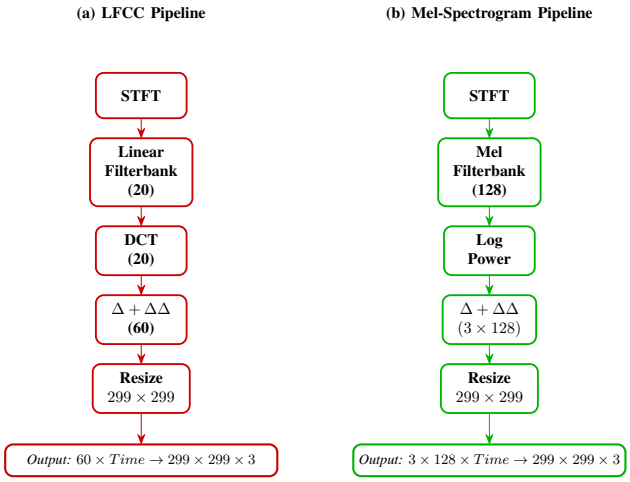

Fig. 3. Block diagram of LFCC and Mel-Spectrogram feature extraction pipelines. The LFCC branch applies linear filtering and a discrete cosine transform, while the Mel-Spectrogram branch preserves dense spectral information prior to CNN input.

C. Model Architecture and Training

We employ Xception architecture (23M parameters) with depthwise separable convolutions [8]. Transfer learning: Phase 1 trains only final layer, Phase 2 fine-tunes entire network. Training: Adam ($lr = 1e^{-4}$), cross-entropy with label smoothing (0.1), batch=128, mixed precision (FP16).

D. Adversarial Robustness

To evaluate detection robustness against evasion atts, we test two gradient-based white-box attacks. For the Fast Gradient Sign Method (FGSM), adversarial perturbations are generated as $\delta = \epsilon \cdot \text{sign}(\nabla_x L)$, where δ denotes the input perturbation, ϵ controls the perturbation magnitude, x is the input feature representation, and L is the classification loss function, with $\epsilon \in \{0.0005, 0.00075, 0.001\}$. For Projected Gradient Descent (PGD), adversarial examples are computed iteratively over 10 steps, where α denotes the step size defined as $\alpha = \epsilon/5$ and $\epsilon \in \{0.1, 0.15, 0.20\}$ [13], [14].

To improve robustness, we implement adaptive adversarial training with dynamic attack sampling weighted by current success rates. Training uses momentum-based updates, where $m = 0.2$ denotes the momentum coefficient, maintains one-third clean samples, where $p = 1/3$ is the probability of sampling clean inputs, to preserve baseline accuracy, and fine-tunes for 10 epochs with a learning rate of 5×10^{-5}.

IV. Experiments and Results

A. Per-Vocoder Analysis

Fig. 4 compares Xception-LFCC (red) and Xception-Mel (blue) across seven neural vocoders. Despite distinct synthesis characteristics and varying detection difficulty, Mel-Spectrograms consistently outperform LFCC, with improvements ranging from 33% to 35%.

Parallel WaveGAN (PWG) is the easiest to detect, with Xception-Mel achieving 0.031 EER (35% improvement over 0.048 EER), attributed to phase discontinuities introduced by parallel generation. WaveGlow follows at 0.033 EER (34% improvement), where flow-based synthesis yields excessive spectral smoothness and implausible temporal coherence.

Multi-Band MelGAN (MB-MelGAN, 0.034 EER) and Full-Band MelGAN (FB-MelGAN, 0.036 EER) show intermediate difficulty with 33% gains, as sub-band decomposition introduces localized boundary artifacts better captured by Mel-Spectrograms' dense frequency resolution (128 bins) compared to LFCC's 20 coefficients.

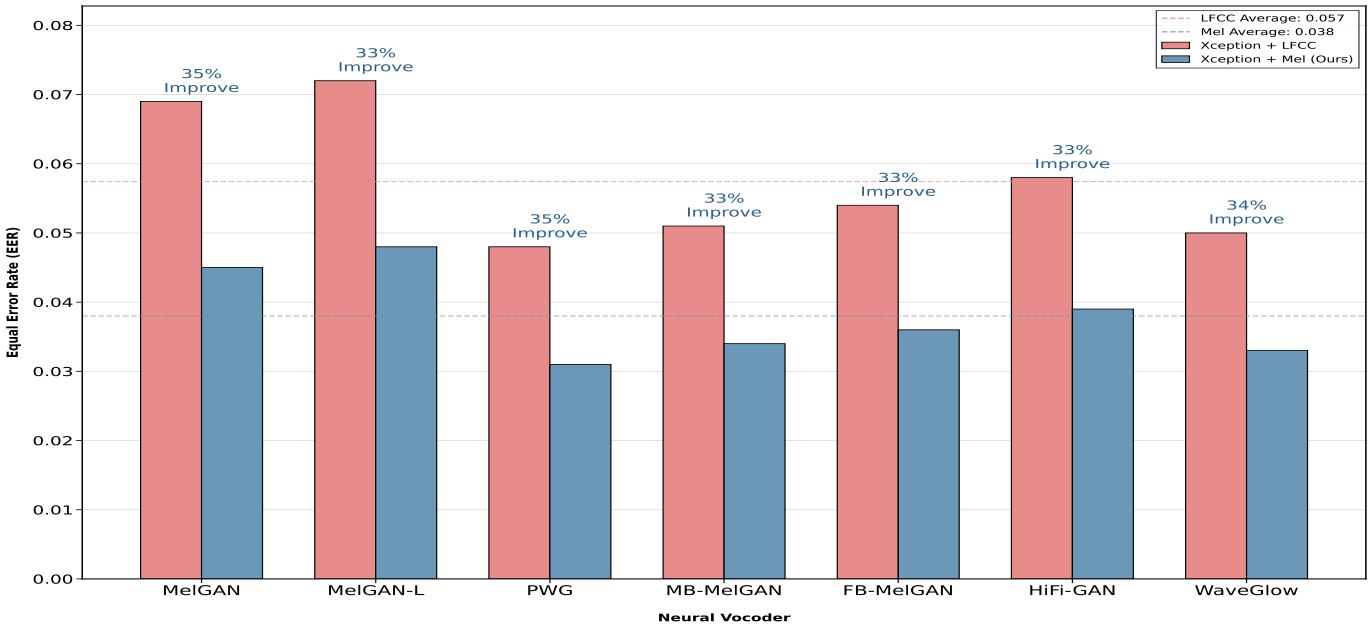

Fig. 4. Per-vocoder leave-one-out cross-validation results

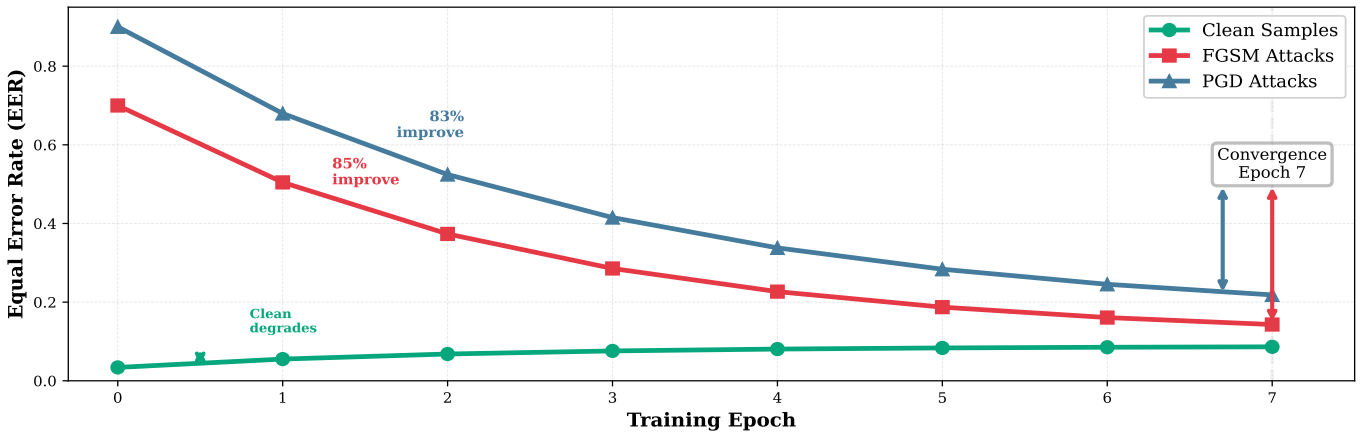

Fig. 5. Convergence of adaptive adversarial training over seven epochs, showing improved robustness against FGSM and PGD attacks and a gradual degradation in clean-sample performance.

HiFi-GAN achieves 0.039 EER (33% improvement), where Mel-based detection identifies subtle pitch micro-variations and unnatural formant transitions during prosodic changes despite advanced adversarial synthesis.

MelGAN and MelGAN-Large present the greatest challenge at 0.045 EER (35%) and 0.048 EER (33%), respectively, reflecting relatively natural global spectra with subtle high-frequency irregularities.

The consistent 33–35% gap across all vocoders indicates that Mel-Spectrograms offer fundamental representational advantages, enabling CNNs to learn synthesis-invariant patterns that generalize across GAN-based, flow-based, and sub-band architectures.

B. Adversarial Training Convergence

Adaptive adversarial training converged within seven epochs, as indicated by stabilization of clean-sample and adversarial EER. Under FGSM, EER decreased from 0.70 to 0.107 (85% improvement), while PGD robustness improved from 0.90 to 0.154 (83% improvement), with a moderate increase in clean-sample EER from 0.034 to 0.088. Fig. 5 shows the training dynamics.

Dynamic attack sampling updates the probability of selecting FGSM or PGD at each epoch based on prior misclassification rates, with higher-success attacks sampled more frequently. Sampling is smoothed using a momentum-based update ($m = 0.2$), and a fixed fraction of clean samples is retained per minibatch to preserve baseline accuracy, yielding rapid robustness gains alongside a gradual degradation in clean-sample performance, reflecting the accuracy-robustness tradeoff.

C. Detection Performance Comparison

Table I presents the Leave-One-Out cross-validation results on the WaveFake dataset, comparing classical baselines with our proposed Mel-Spectrogram-based approach. Experiment I represents as Xception with LFCC feature and Experiment II as Xception with Mel Spectrogram, respectively.

TABLE I
LEAVE-ONE-OUT CROSS-VALIDATION RESULTS ON THE WAVEFAKE DATASET

Method	Feature	Params	Avg EER	Best	Worst
GMM [7]	LFCC	∼10K	0.058	0.013	0.220
RawNet2 [9]	Raw	15M	0.040	—	—
Experiment I	LFCC	23M	0.053	0.042	0.068
Experiment II	**Mel**	**23M**	**0.034**	**0.028**	**0.041**
Improvement	—	—	**36%**	**33%**	**40%**

D. Adversarial Robustness Results

Table II demonstrates that large-scale pretrained CNNs achieve substantial adversarial robustness through adaptive training. Our Xception-Mel model with adaptive adversarial training reduces FGSM attack error rate by 86% (from 0.70 to 0.107 EER) and PGD attack error rate by 83% (from 0.90 to 0.154 EER). Despite fifty times more parameters than the lightweight LCNN model [10], our approach achieves comparable improvement percentages: 86% for FGSM and 83% for PGD, suggesting that defense effectiveness scales across model capacities. Clean performance degradation (0.034 to 0.088 EER) represents an acceptable accuracy-robustness tradeoff, remaining competitive with classical GMM baselines (0.058 EER) [7] while providing substantial adversarial robustness unavailable in statistical models. We made our source code available at https://github.com/NinjaIfti/Thesis, where Experiment II implements Xception with Mel-Spectrograms and Experiment III implements adaptive adversarial training.

TABLE II
ADVERSARIAL ROBUSTNESS: STANDARD VS ADAPTIVE TRAINING

Model	Clean EER	FGSM EER	PGD EER	FGSM Improve	PGD Improve
Experiment II	0.034	0.70	0.90	—	—
Experiment III	**0.088**	**0.107**	**0.154**	**86%**	**83%**

V. Discussion

A. Mel-Spectrograms Outperform LFCC for CNNs

Our results show that optimal feature representation depends on classifier architecture. LFCC preserves high-frequency vocoder artifacts on a linear scale, explaining its strong performance with Gaussian Mixture Models [7]. In contrast, Convolutional Neural Networks benefit from dense two-dimensional inputs. Mel-Spectrograms provide 128 frequency bins compared to LFCC's 20 cepstral coefficients, offering 6.4× greater spectral detail and richer spatial structure for convolutional filters. Their grid-like representation also enables effective transfer learning from ImageNet-pretrained models, whereas LFCC's sparse one-dimensional form limits spatial pattern learning and transfer efficiency.

This performance gap extends to cross-vocoder generalization. Leave-one-out evaluation shows that Mel-Spectrogram models maintain strong performance on unseen generators, while LFCC models degrade substantially. This indicates that CNNs learn synthesis-invariant features from dense spectral representations rather than memorizing generator-specific artifacts from compressed cepstral features.

B. Practical Implications

These findings guide feature selection for audio deepfake detection. Dense spectral representations are well suited to CNN-based detectors, while sparse cepstral features favor classical statistical models. For CNN deployments, Mel-Spectrograms should be the default choice due to advantages in accuracy, generalization, and baseline adversarial robustness.

Adaptive adversarial training enables robust CNN-based detectors. The accuracy-robustness tradeoff (clean EER increase from 0.034 to 0.088 for 85–83% attack improvement) represents an acceptable compromise for security-critical applications. However, adversarial training approximately doubles training time, and this overhead must be balanced against deployment constraints.

C. Limitations and Future Work

First, evaluation is limited to the WaveFake dataset with a single female speaker recorded under studio conditions. Cross-dataset validation on ASVspoof, FakeAVCeleb, and real-world audio would strengthen generalization claims [6], [15].

Second, our analysis focuses on white-box gradient-based attacks. More advanced adaptive, perceptual, and physical-world attacks, as well as black-box and gray-box threat models, warrant investigation to provide realistic deployment guarantees.

Third, computational efficiency remains a challenge. Although real-time inference is achieved, adversarial training limits rapid updates. Future work should explore curriculum learning, knowledge distillation, and early stopping to reduce training cost while preserving robustness.

Finally, feature design remains open. Beyond dense Mel-Spectrograms and sparse LFCC, intermediate or hybrid representations and learned features via end-to-end optimization or neural architecture search may yield improved accuracy-efficiency-robustness tradeoffs.

VI. Conclusion

This work demonstrates that feature representation must align with detector architecture. While LFCC remains effective for lightweight statistical models, dense spectral representations such as Mel-Spectrograms better exploit CNN capacity, enable transfer learning, and improve adversarial robustness. Future directions include cross-dataset generalization, computationally efficient adversarial training, and adaptive detection frameworks that evolve with emerging synthesis methods.

References

[1] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[2] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 14 910–14 921.

[3] R. J. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53145796

[4] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[5] J. Damiani. (2019) A voice deepfake was used to scam a CEO out of $243,000. [Online]. Available: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[7] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[9] J. weon Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech 2022*, 2022, pp. 2438–2442.

[10] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio deepfake detection," in *Proc. Interspeech 2023*, 2023, pp. 5276–5280.

[11] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR 2015*, 2015.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR 2018*, 2018.

[15] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.