GGMDDC: A GAN-Guided Multilingual Audio Deepfake Detection Dataset

Ravindrakumar M. Purohit, Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202321002, 202101154, hemant_patil}@daiict.ac.in

***Abstract*—Audio Deepfake Detection (ADD) is of critical concern due to its social relevance. The development of ADD solution faces the limitation of the low diversity of language resources and is restricted to a limited number of speakers. However, there is a need for a more reliable ADD system due to advancements in deepfake audio generation by the attackers. Most of the existing datasets for ADD task are limited to a language count, so their usage remains limited to restricted number of speakers. Also, in most of the existing datasets for ADD, the number of issue of some speakers' real and deepfake exists, and others are unavailable. In this study, we implemented a configuring version of Generative Adversarial Networks (GANs), namely, High Fidelity-GAN, to create dataset, GAN-Guideded Multilingual Deepfake Detection Corpus(GGMDDC). We achieved Mean Openion Score (MOS) of 4.52 for the proposed system, which proved to be better than MOS of the existing systems.**

***Index Terms* - HiFi-GAN, Mel Spectrogram, Multilingual Deepfakes, Audio Deepfake Detection.**

I. Introduction

Deepfake is the synthetic data generated, converted, or edited using the Deep Learning (DL) algorithm with the goal of making fake audio, video, or image that sounds real. In recent years, deepfake content in image, video, and audio formats has become a boom with advancement in technology. In the earlier days, it took a lot of work to generate deepfake because of the complex nature of data patterns or the limited computational resources.

Deepfake signals are generated by preserving the voice and sentence structure, in order to make them close to real speech signals. Innovation of Generative Adversarial Networks (GANs) in 2014, resulted in a solution to many problems e.g., data augmentation using generative models [1], [2]. The potential of GAN for ADD task has also been explored previously [3]. It is motivated by recent advancements in GAN, particularly, High-Fidelity Generative Adversarial Networks (HiFi-GAN), to generate deepfake audio from real utterances. HiFi-GAN (High fidelity means, closely matches the natural human voice in terms of clarity, accuracy, and realism) works in two phases. In the first phase, it predicts the low resolution intermediate representation of linguistic features in terms of the Mel spectrogram [4], and in the last phase, it creates the synthetic raw waveform from the intermediate representation of the Mel spectrogram [5]. During this process, we took sampling frequency as 22050 samples per second, all representing the 16-bits fidelity. As output, raw audio will closely resemble the qualities of human speech. We created a HiFi-GAN model to generate the dataset. At first, we trained a model using the LJ Speech dataset, in which we employed a generator (G), and discriminator (D). We trained the model with the LJ Speech audio dataset containing 13,100 samples. After 600k steps, we got an optimized model, which was further used to generate deepfake files. For inference purposes, the samples pass through the upsampling (22050 Hz from 16000 Hz) phase, where a pre-trained feature extraction module extracts Mel spectrograms, which are further passed through a generator to extract and learn the speaker's properties. Further, it will pass through a discriminator to test the quality of the generated signal.

A. Related Work

Many studies have been lately focusing on Audio Deepfake Detection (ADD) task [7]. However, attackers succeed to attack many times due to the perceptually similar creation of deepfake audio, which creates threat to audio insecurity. Many researchers have started working on ADD tasks, and many datasets have been developed to safeguard society against the adverse effects of deepfakes. However, most of the existing datasets are restricted to one language (English) speakers only, resulting in difficulty to detect deepfakes in another language. One of the most popular existing dataset for ADD task, namely, FoR (Fake or Real) is also restricted to English dialect only [8]. Other popular datasets, namely, in-the-wild (ITW) [9], Singfake [10], and ASVSpoof [11], [12] datasets are also released with English and Mandarin language speakers recordings only.

Recently released dataset, namely, WaveFake [13], contains audio samples in two languages (English and Japanese). Many a times, attackers plan to generate deepfake audio of celebrities, politicians, or businessman in native languages, resulting in a failure of good performance models trained on one language only. In highly populated countries, such as China, and India, native language-based deepfakes seems as real to locals, creating a threat to subjects' personality, security, and business. In this study, we propose a dataset (containing samples collected from 10 different languages), namely, GAN-guideded Multilingual Corpus for Deepfake Detection Corpus (GGMDDC). This dataset is aimed to boost ADD research activity, due to its property of being robust to speakers' dialect. It contains some of the most popular languages of world, such as Russian, French, Hindi, etc. resulting in its broad application perspective for ADD task.

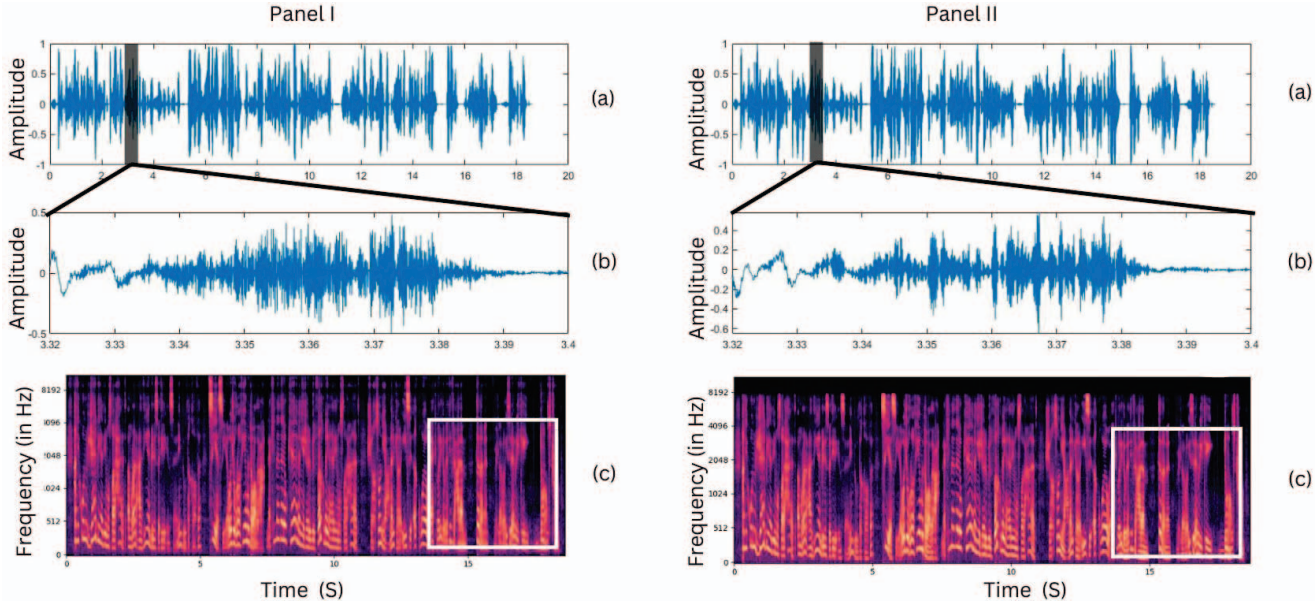

Fig. 1. Panel I (deepfake signal), and Panel II (real speech): (a), (b), and (c) shows the waveform, short speech segment, and Mel spectrogram, respectively (White boxes in Panel I (c) and Panel II (c) highlight the similarities between the deepfake and real speech samples).

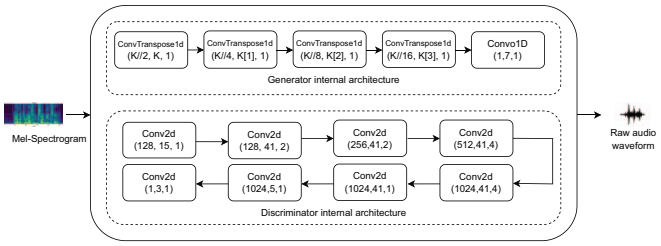

Fig. 2. Architecture of HiFi-GAN. Adapted from [6].

Fig. 1 (a) shows that mostly both signals (deepfake as well as real) are similar to each other. However, a significant difference can be noted in the amplitude variations in both waveforms. It can be observed that the deepfake signal has more high amplitude samples than a low number of high amplitude samples. Also, in Fig. 1, it can be observed that Panel I and Panel II show the variational differences in amplitude over time. In the generated speech spectrogram, there are three bands with a broader frequency range compared to Panel II, or real audio. The real audio displays the thin and well-defined bands, indirectly showing the narrower frequency range as more focused and natural speech. The distribution of the frequency and amplitude over time helps identify the deepfake audio from the spectrogram characteristics.

II. Proposed Methodology

A. Why HiFi-GANs?

In the case of deepfake detection, the dataset should contain high quality and controllable speech samples. To create GGMDDC, we employ HiFi-GAN speech synthesis architecture, to create deepfake audio samples from the real datasets. To choose the appropriate architecture, we surveyed several architectures, such as Tacotron [14], Tacotron2 [15], Deep Voice 3 [16], clariNet [17], Parallel WaveGAN [18], and many more (as mentioned in Table I). Tacotron is an end-to-end TTS model, which follows the *seq2seq* paradism to synthesis the speech [14]. Tacotron predicts the Mel spectrogram from the text and later on uses the Griffin-Lim reconstruction algorithm [19] to construct the raw waveforms. However, model struggle with unusual or unexpected text inputs, resulting in incorrect or nonsensical outputs, and achived the MOS of 3.82 out of the *5*-scale. The Tacotron2 [15] uses the recurrent *seq2seq* network with an attention mechanism to predict the Mel spectrograms, and later on vocoder transforms the Mel spectrogram to the raw waveform. Deep Voice 3 [16], follows the fully-convolutional architecture with monotonic attention mechanism, to minimise the limitation of the *seq2seq* models for waveform synthesis, again, the control over the prosody is the less as compared to the Tacotron2 and achieves the MOS of the 3.62, which was less than the first version of the Tacotron. ClariNet [17] and Parallel WaveGAN [18] are limited to single-stage speech synthesis models. Where ClariNet [17] is built on an autoregressive (AR) [18] Text-to-Wave (T2W) neural network, and Parallel WaveGAN employs a non-autoregressive (NAR) WaveNet architecture. ClariNet provides faster inference speed because of its AR structure. However, because of its fully parallel design, Parallel WaveGAN [18] trains far more quickly than ClariNet [17]. In particular, it takes just 2.8 days compared to the 12.7 days for ClariNet [17]. Despite this discrepancy, both models, i.e., ClariNet [17] [17] and parallel WaveGAN [18] achieve most of 4.15 and 4.16, respectively, and achieve comparable quality while synthesizing raw waveforms. One of the major issues in almost every existing work is Real-Time Capability (RTC), which is provided by a few models. The RTC models have low MOS compared to HiFiGAN, forcing us to employ HiFi-GAN for deepfake generation. As

shown in Table I, Transformer TTS [20], HiFiGAN [6], FreGAN [21], Fastspeech2 [22], Glow-TTS [23], and BigVGAN [24]. All models perform a significant amount of the speech synthesis quality but in terms of the other parameters, e.g., Quality-of-Speech (QoS), inference speed, real-time capability, computational cost, and multiple modalities. Recent models have significantly improved upon previous works, bringing the quality of synthesized audio closer to natural human speech. However, many models, such as Tacotron [14], lack the RTC or require significant computations for fast inference. In contrast, HiFi-GAN performs high inference speed without generating poor-quality samples [6]. In such cases, handling multi-speakers and diverse speech modalities is essential to generate robust deepfake audio, which tends to struggle for AR and NAR when a dataset containing multiple speakers or different speech styles. To balance the quality, speed, and robustness against various inputs, make the HiFiGAN a superior choice over other architectures like Tacotron [14], clariNET [17], and parallel WaveGAN [5].

Additionally, they have greatly reduced training and inference times without degrading speech synthesis performance. We initially started with the LJ Speech dataset, which includes the speech of a single female speaker for model training. However, despite its high quality recordings, the dataset lacks speaker diversity, and the result could have been more optimal. Consequently, we fine-tuned the model on the VCTK dataset (which contains recordings from 109 speakers with various accents and speaking styles) using transfer learning. This approach significantly improved the ability of the model to generate deepfake audio that closely resembles the original speech samples in the real dataset. Fig. 1 illustrates the high degree of similarity achieved between the original and synthetic samples.

The generator architecture [6] is initially built using the convolutional neural networks as shown in Fig. 2. It starts with the convolution layers (ConvTranspose1D) and weight normalization (each layer followed by a nonlinear activation function $f(\cdot)$), which helps to extract the different aspects of the Mel spectrogram from the input features by decoupling the magnitude of the weight vector from the Mel spectrogram, also it progressively increases the temporal resolutions of the signal, and do feature mapping-stretches over time according to the specific upsample rates for the high frequency details. The feature maps are refined to the final audio waveform with the single output channel at the final layer by normalizing the weights. However, the MRF (Multiple Receptive Field-fusion) [26] block is also used in generator architecture because it identifies the various lengths of hidden patterns from the Mel spectrograms. A large receptive field helps to identify the phonemes more effectively [4], [26]. Resblock1 and Resblock2 are multi-scale architectures that apply multiple convolutional operations to the different delations rated to capture *local* and *global* temporal dependencies from nonlinear transformations and residual connections. The feature maps are refined to the final audio waveform with the single output channel at the final layer by normalizing the weights.

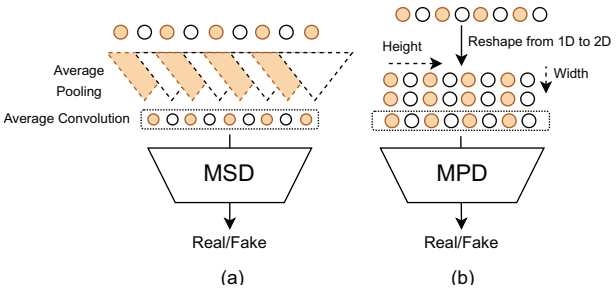

Fig. 3. Architecture of discriminator. After [6].

1) Discriminator: The discriminator D identifies the short and long-term dependencies on the raw waveform made by the generator. To achieve natural-sounding speech synthesis, it is crucial to understand the long-term dependencies in speech signal. This ensures that the generated speech closely mimics natural human speech. Studies reported in [4], [14] treat audio as a sinusoidal signal with varying periods to capture the nonlinear correlations between different phonemes.

To capture various repeating patterns, the discriminator (refer Fig. 3) is divided into two sub-discriminators: **(1) Multi-Period Discriminator (MPD), and (2) Multi-Scale Discriminator (MSD)**. D_{MPD} is a mixture of multiple sub-discriminators. It tries to capture the implicit features of the raw waveform with a different period sizes to avoid overlapping from a length, $\boldsymbol{T}$. Mathematical representation as follows in Eq. (1), where $\boldsymbol{N}$ represents the number of sub-discriminators, $\boldsymbol{D}$ denotes the contribution of each sub-discriminator to the overall discriminator's decision, i.e.,

$$D_{MPD} = \sum_{i=1}^{N} D_i. \tag{1}$$

D_{MSD} architecture was proposed first by Kumar *et al.* [26]. MSD combines the three sub-discriminators, e.g., raw audio, *2 average-pooled audio, and *4 average-pooled audio. Each of the discriminator convolutions is covered using the leaky ReLU. Discriminator uses the *weight* and *spectral* normalization [27] to do stable training, and avoid exploding gradient problems in discriminator networks.

2) Loss Functions: As the original papers of the GANs [28], [29] utilize minimax as shown in Eq. (2) and Wasserstein loss functions, operating on distance or probability distributions between generated and real data. Minimax loss establishes a competitive game for realistic sample generation. Wasserstein loss in WGANs focuses on minimizing Wasserstein distance for smoother gradients. In particular,

$$\max_{D} V(D) = E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))]. \tag{2}$$

Loss function makes a balance between the generator and discriminator. WGAN refines Wasserstein loss with a gradient penalty, enhancing stability in GAN training. The loss function of both networks is next.

1) **Generator Loss Function**: Let G denote the generator, D denote the discriminator, and α be a reward/penalty

TABLE I
COMPARISON OF VARIOUS ARCHITECTURES BASED ON VARIOUS EVALUATION FACTORS

Architecture	Dataset	5-Scale MOS	QoS (MOS)	Inference Speed	RTC[$]	Multiple Modalities	Computational Cost	Limitations
Tacotron [14]	NAE	3.82±0.08*	Fair	Moderate	No	Moderate	Moderate	Alignment issues and artifacts issues generates mechanical-sounding voice.
Tacotron 2 [15]	LJ Speech	4.52±0.06*	Good	Moderate	No	Good	High	Slower inference speed and higher computational cost.
Deep Voice 3 [16]	VCTK	3.01±0.29*	Fair	Moderate	No	Good	High	Higher computational requirements.
ClariNet [17]	IES	4.15±0.25*	Good	Moderate	No	Excellent	Moderate	Training can be complex and resource-intensive.
Parallel WaveGAN [18]	Custom Dataset	4.16±0.09*	Good	Very Fast	Yes	Good	Low	May not capture fine speech details as well as others.
Transformer TTS [20]	US English female	4.44±0.05*	Good	Moderate	No	Good	Moderate	Requires significant training data.
HiFi-GAN [25]	**LJSpeech**	**4.36±0.07***	**Good**	**1,186×**	**Yes**	**Excellent**	**Moderate**	**Can be resource-intensive in some cases.**
Fre-GAN [21]	LJ Speech	4.25±0.04	Good	Moderate	No	Good	Moderate	Requires careful tuning for best results.
FastSpeech 2 [22]	LJ Speech	3.83±0.08	Fair	Very Fast	Yes	Good	Low	Lower expressiveness compared to some models.
Glow-TTS [23]	LJ Speech	4.01±0.08	Good	Moderate	No	Good	Moderate	More complex training process.
BigVGAN [24]	LibriTTS	4.11±0.09	Good	Fast	Yes	Good	Low	May struggle with very diverse speakers.

* shows confidence interval is 95%. [$] RTC - Real-Time Capability, NAE - North Americal English, QoS = Quality of Speech.

factor. The generator loss (L_G) can be formulated, in simple words as:

$$G_L = [\alpha \cdot \text{reward} + (1-\alpha) \cdot \text{penalty}]. \tag{3}$$

In the other words,

$$G_L = \theta_g \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))). \tag{4}$$

2) **Discriminator Loss Function**: Let G denote the generator, D denote the discriminator, and α be a reward/penalty factor. The discriminator loss (L_D) can be formulated as follows, in simple words:

$$D_L = \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right]. \tag{5}$$

It penalizes itself for misclassifying a real instance as fake or a fake instance (created by the generator) as real by maximizing the Eq. (5).

3) Training Paradigm: The training paradigm of the HiFi-GAN follows the Goodfellow *et al.* [28]. The primary goal of GAN training is to find an equilibrium, where the generator produces realistic data that fool the discriminator. As per Eq. (6):

$$\begin{aligned} \min_G \max_D V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] \\ + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \end{aligned} \tag{6}$$

The training of the HiFi-GAN is being used for speech-related tasks, e.g., ASR, Text-to-Speech (TTS) synthesis [26]. Training is set up using the multiple loss functions described earlier.

III. DATASET FORMATION

This Section proposes the presents dataset structure and its design details. Due to the limited language resources, the authors were unable to collect data manually as real samples were collected for our dataset. Alternatively, in this study, we propose a dataset in which we generated 40,000 deepfake samples of real utterances that were collected from the VoxLingua107 dataset [30], which is one of the most popular largest open source multilingual dataset for Spoken Language Identification (SLID) task. VoxLingua107 was formed by recording utterances from 107 different languages and data from 6628 hours. Limited to storage resources, authors could not create a dataset for more than 10 languages, namely, Russian, French, Arabic, Spanish, Vietnamese, Mandarin Chinese, English, Hindi, Portuguese, and Sanskrit. Around 11.35 hours of data was selected from each language having based on time duration statistics. Comprising 10 languages in the GGMDDC, it is also robust to speaker's dialects. To our best knowledge and belief, this is the first study of its kind that employs and proposes the corpus, which contains both the real and the corresponding fake utterances of each speaker w.r.t. text material used for the recordings. The total number of utterances in the proposed dataset is 80,000 (40,000 real and 40,000 fake), making it one of the largest datasets among currently available open source datasets in the literature. Dataset statistics and demo is publicly available at [1].

A. Real Data

First, we collected all the samples available from the VoxLingua107 dataset (which is an open source freely available dataset) [30]. We labeled them into 5 classes based on the audio duration of particular samples, namely, A (0-5 seconds),

[1]For more details, https://iamshreeji-copy1.github.io/submission_website/

B (5-10 seconds), C (10-15 seconds), D (15-20 seconds), and E ($>$ 20 seconds). After that, we selected 1,000 random samples (except in one (Sanskrit) class) from each class collectively to form a dataset of total of 111.16 hours (40,000 samples) of real data. We eliminated the issue due to audio sample size dependencies by selecting the variable length audio. In order to generalize sampling rate to *16 kHz*, all audios of VoxLingua107 were resampled to 16 kHz before generating deepfake files from it. The resampling process was carried out in order to generalize the dataset.

B. Fake Data

We use the model based on HiFi-GANs to generate deepfakes from the real signal (described in subsection II-A). We employed to process the real audio and generate the deepfake audio of the same speaker with the same utterance spoken in the original sample. HiFi-GAN generated deepfake illustrate almost similar properties as the real signals. Audio-based comparison of generated deepfakes and real, as discussed in sub-Section I-A, illustrates the difficulties in distinguishing between fake *vs.* real. Due to their perfect creation (i.e., high *perceptual* similarity), these generated deepfakes are extremely difficult to distinguish between human ears. As the dataset is generated by sophisticated machine learning/DL methods, it also aims to fool the ADD system and humans.

C. Results

Table II compares the performance of various datasets using several metrics, namely *Perceptual Evaluation of Speech Quality* (PESQ), *Short-Time Objective Intelligibility* (STOI), *Fréchet Distance, Mean Opinion Score* (MOS), *Subjective Mean Opinion Score* (SMOS), and *Mel Cepstral Distortion* (MCD). The proposed dataset demonstrates superior performance across these metrics. Specifically, it achieves a PESQ score of 1.0213, which is competitive with the other datasets. The STOI score of 0.1224 is notably higher, indicating better intelligibility. Additionally, the proposed dataset records the lowest Frechet distance of 220.22, suggesting it generates deepfakes that are perceptually closest to the real audio. The Mean Opinion Score (MOS) of 4.52, and Subjective Mean Opinion Score (SMOS) of 4.39 are the highest among all datasets, reflecting higher quality and naturalness of the generated audio. Lastly, the Mel Cepstral Distortion (MCD) score of 23.638 is the lowest, indicating better spectral quality. These results collectively demonstrate that the proposed dataset, when used with HiFi-GAN, is capable of generating more convincing and high quality deepfakes compared to the existing datasets.

TABLE II
COMPARISON W.R.T. VARIOUS EXISTING DATASETS

Dataset	PESQ	STOI	Frechet Distance	MCD	MOS	SMOS
FoR [8]	1.02	-0.01	228.74	4.12	4.02	24.95
ITW [9]	1.06	0.22	287.66	34.20	3.64	3.46
Singfake [10]	1.32	0.07	324.31	34.22	4.36	4.38
Wavfake [13]	**2.91**	0.95	276.3	34.82	4.4	4.11
Proposed	1.02	**0.12**	**220.22**	**23.63**	**4.52**	**4.39**

IV. SUMMARY AND CONCLUSIONS

In this study, we proposed GGMDDC dataset, which is generated using recently proposed, HiFi-GAN. Real data in GGMDDC dataset was acquired from VoxLingua107 dataset, whose deepfakes were generated GGMDDC. Total of 40,000 files collectively form GGMDDC dataset. A proposed dataset is a multilingual dataset, which consists of 10 different languages, resulting into a wide application of dataset. We also proposed a few experiments on GGMDDC dataset, resulting into a comparable accuracy. Future works involve more detailed experimentation and analysis on proposed dataset using advanced DL methods. One key limitation of this study would be the GAN architecture used in this work use pretrained models in English to generate multilingual voices in 10 different languages. Limitations of this study include training GANs on the LJ Speech and VCTK dataset, i.e., a single language dataset. Capturing the critical phoneme patterns of different languages is challenging when using a dataset from just one language. In the future, We aim to address this limitation by training the model on diverse languages and large-scale datasets to overcome this issue. However, the model's effectiveness will increase when trained on the multilingual dataset and provides the generalized diverse and dialects linguistic characteristics to identify the deepfake. The rapid involvement of the deepfake generation is already a challenge for deepfake detection. Implementing the robust multilingual dataset is computationally intensive due to the detection in diverse scenarios in real-time, which makes ongoing research on ADD more complicated and challenging, which remains an open research challenge associated with our detection process.

REFERENCES

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017, {Last Accessed Date : 1^{st} August, 2024}.

[2] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, 106–110, Brighton, United Kingdom.

[3] T. P. Doan, K. Hong, and S. Jung, "GAN discriminator based audio deepfake detection," in *Proceedings of the 2^{nd} Workshop on Security Implications of Deepfakes and Cheapfakes*, 2023, Melbourne, Australia, pp. 29–32.

[4] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016, {Last Accessed Date : 1^{st} August, 2024}.

[5] A. Oord, Y. Li, I. Babuschkin, *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning (ICML)*, PMLR, 2018, Stockholm, Sweden, pp. 3918–3926.

[6] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 17 022–17 033, 2020, {Last Accessed Date : 1^{st} August, 2024}.

[7] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, vol. 2023, 2023, 2808–2812, Dublin, Ireland.

[8] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, Timisoara, Romania, pp. 1–10.

[9] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022, {Last Accessed Date : 9^{th} March, 2024}.

[10] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," *arXiv preprint arXiv:2309.07525*, 2023 {Last Accessed Date : 9^{th} March, 2024}.

[11] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[12] A. Chintha, B. Thai, S. J. Sohrawardi, *et al.*, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[13] J. Frank and L. Schönherr, "Wavefake: A dataset to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021 {Last Accessed Date : 9^{th} March, 2024}.

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017 {Last Accessed Date : 1^{st} August, 2024}.

[15] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, Calgary, Canada, pp. 4779–4783.

[16] W. Ping, K. Peng, A. Gibiansky, *et al.*, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017 {Last Accessed Date : 1^{st} August, 2024}.

[17] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018 {Last Accessed Date : 1^{st} August, 2024}.

[18] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, Virtual Barcelona, pp. 6199–6203.

[19] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, {Last Accessed Date : 1^{st} August, 2024}, pp. 6706–6713.

[21] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial frequency-consistent audio synthesis," *arXiv preprint arXiv:2106.02297*, 2021, {Last Accessed Date : 1^{st} August, 2024}.

[22] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020 {Last Accessed Date : 1^{st} August, 2024}.

[23] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 8067–8077, 2020, Vancouver, Canada.

[24] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022 {Last Accessed Date : 1^{st} August, 2024}.

[25] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020, {Last Accessed Date : 1^{st} August, 2024}.

[26] K. Kumar, R. Kumar, T. De Boissiere, *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 14910-14921, 2019, Vancouver, Canada.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018 {Last Accessed Date : 1^{st} August, 2024}.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Neural Information Processing Systems (NIPS)*, vol. 27, pp. 2672-2680, 2014, Montreal, Canada.

[29] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016 {Last Accessed Date : 1^{st} August, 2024}.

[30] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, Shenzhen, China, pp. 652–658.