Diffusion-GAN Model for Audio Synthesis

Suhwan Sung*, Seongho Keum*, Hyeongseok Gwak*, Sungwoo Jeon*
Korea Advanced Institute of Science and Technology
{acgnkpiyrw, keum07, khs5696, j0070ak}@kaist.ac.kr

Abstract

Audio synthesis is a fundamental task in the field of artificial intelligence and signal processing, with applications ranging from speech synthesis to music generation. In this paper, we propose a novel approach called the Diffusion-GAN model for audio synthesis. We use the Diffusion-GAN model, which combines the principles of diffusion models and generative adversarial networks (GANs), to generate high-quality and realistic audio samples. We conduct experiments using the Speech Command Zero through Nine dataset, demonstrating the efficacy of the Diffusion-GAN model and its superiority over existing approaches. Our results showcase the potential of the proposed model for various audio synthesis applications, including speech synthesis, music generation, and sound design. Further investigation into evaluation criteria, audio embedding methods, alternative base models, and parameter optimization is recommended to enhance the capabilities and performance of the Diffusion-GAN model. Overall, our research contributes to the advancement of audio synthesis techniques and provides a valuable benchmark for future studies in this domain.

1. Introduction

Audio synthesis, the process of generating realistic and high-quality audio signals, has been an active area of research with numerous applications in fields such as music production [2] [8], speech synthesis [1], and sound design [5]. The development of effective models for audio synthesis is crucial for producing authentic and diverse audio content.

In recent years, diffusion models and generative adversarial networks (GANs) have emerged as powerful techniques for generating realistic and complex data, including images, text, and audio. Diffusion models, in particular, aim to produce output samples with a probability distribution similar to that of the input data by iteratively adding and removing noise. DDPM [4] and many recent works showed that the diffusion method is a strong and powerful add-on method for neural networks. This process, known as forward diffusion and reverse denoising, enables the generation of high-fidelity samples that capture the underlying structure of the data.

In the domain of audio synthesis, notable advancements have been made by researchers in applying diffusion concepts to generate realistic voices and achieve consistent pronunciation. DiffWave [7], for instance, has successfully employed diffusion-embedding techniques to produce convincing voices in both unconditional and class-conditional settings. HiFi-GAN [6], on the other hand, has utilized GAN architectures with multiple discriminators to achieve higher computational efficiency and improved sample quality.

While diffusion models and GANs have individually demonstrated their effectiveness in audio synthesis, the combination of these two approaches, specifically Diffusion-GAN [9], remains relatively unexplored. By integrating the diffusion process into the GAN framework, there is potential for improved data efficiency, stability in model training associated with GANs.

In this paper, we propose a novel method, *Diffusion-GAN for audio synthesis*, which leverages the strengths of diffusion models and GANs to generate high-quality audio samples. We aim to fill the gap in the existing literature by applying the diffusion-GAN approach to audio synthesis, demonstrating its advantages over traditional diffusion models and GAN-based audio generators.

To evaluate the performance of our proposed method, we employ the Speech Command Zero through Nine dataset [10], a widely-used dataset containing digit pronunciations collected from numerous speakers. By training the Diffusion-GAN model on this dataset, we demonstrate its ability to generate realistic audio samples and compare its performance with other existing models, such as DiffWave.

The contributions of this research include the development and evaluation of the Diffusion-GAN model for audio synthesis, providing insights into the benefits and limitations of the approach. In the following sections of this

[0]* Equal contribution.

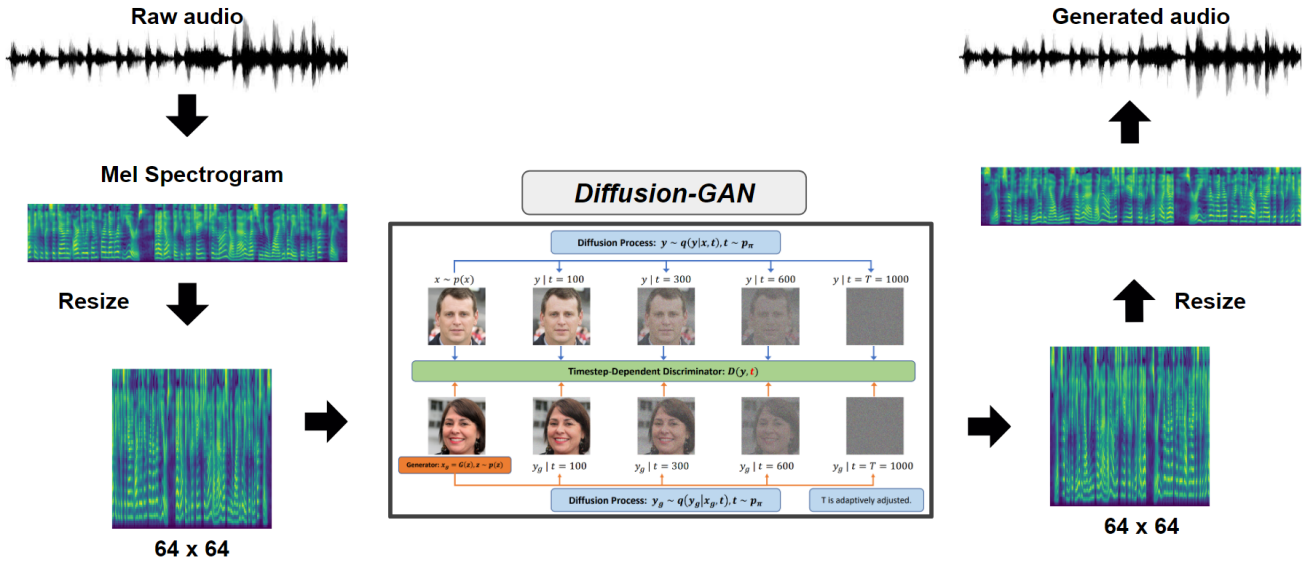


Figure 1. Diagram for overall architecture.

report, we will delve into the related work, describe the methodology behind the Diffusion-GAN model, present experimental results, and conclude with a comprehensive discussion of the findings and their implications.

2. Related Work

In the field of audio synthesis, several notable studies have contributed to the advancement of diffusion models and GANs. Understanding the existing research is crucial for contextualizing the novelty and significance of the proposed Diffusion-GAN model for audio synthesis.

2.1. DiffWave

DiffWave [7], a recent research work, stands out as a successful application of the diffusion concept to the field of audio synthesis. DiffWave focuses on generating realistic voices with consistent word-level pronunciation in both unconditional and class-conditional settings. The model employs a structure where audio inputs and embedded vectors undergo the diffusion process, yielding high-quality outputs. The utilization of diffusion-embedding in DiffWave has demonstrated impressive performance in generating audio samples with excellent fidelity.

2.2. HiFi-GAN

Another notable advancement in audio synthesis is HiFi-GAN [6], a GAN-based audio generator that achieves both higher computational efficiency and sample quality. HiFi-GAN comprises a generator and two discriminators. The multi-period discriminator (MPD) identifies diverse periodic signals underlying the audio data, while the multi-scale discriminator (MSD) focuses on detecting consecutive patterns and capturing long-term dependencies. This multi-discriminator architecture has proven effective in improving both the computational efficiency and the quality of generated samples. The generator in HiFi-GAN takes the input mel-spectrogram and employs transposed convolution and the multi-receptive field fusion (MRF) layer for upsampling, ensuring that the resolution of the generated raw waveforms matches that of the input spectrogram.

It is worth noting that while diffusion models and GANs have demonstrated individual successes in audio synthesis, their combination in the form of diffusion-GAN remains relatively unexplored. While several studies have developed diffusion models and GANs independently, few have explored their integration. Therefore, the application of the diffusion-GAN approach to audio synthesis presents a unique opportunity to leverage the advantages of both diffusion models and GANs, potentially leading to improved performance and novel insights.

3. Method

This section outlines the methodology employed in the proposed Diffusion-GAN model for audio synthesis. It describes the overall architecture, training procedure, and key components of the model.

3.1. Architecture

The Diffusion-GAN model for audio synthesis combines the principles of diffusion models and GANs to generate high-quality audio samples. The architecture1 consists of a generator and a discriminator, which are trained in an adversarial manner.

The generator takes as input a mel-spectrogram, which captures the frequency content of the audio signal. The sequence length of the generated waveforms is adjusted to match the resolution of the original audio data.

The discriminator, on the other hand, aims to distinguish between real and generated mel-spectrogram samples. It receives the generated mel-spectrogram and the corresponding time step of the diffusion process as inputs. The discriminator is trained to provide a higher score for real mel-spectrograms and a lower score for generated ones.

3.2. Training Procedure

The training of the Diffusion-GAN model follows an adversarial learning framework. The generator and discriminator are trained iteratively, competing against each other to improve the quality of the generated mel-spectrograms.

During training, the generator takes mel-spectrograms as input and generates corresponding raw waveforms. These generated samples are then passed to the discriminator, along with the corresponding time step of the diffusion process. The discriminator provides feedback by assigning scores indicating the authenticity of the samples.

The training objective is to minimize the discriminator's ability to distinguish between real and generated mel-spectrograms, while simultaneously maximizing the generator's ability to produce realistic and high-quality samples. This adversarial training process encourages the generator to improve its output distribution to resemble that of real mel-spectrogram.

3.3. Data Preparation and Preprocessing

To facilitate training and evaluation, the audio data used in the Diffusion-GAN model is typically preprocessed. Raw audio data is segmented into smaller segments, typically at a granularity of 1 millisecond, to capture temporal variations in the audio signal. Each segment is then transformed into a 2D representation known as a mel-spectrogram, which represents the frequency content of the audio signal. The use of mel-spectrograms allows for efficient processing and reduces the dimensionality of the input data.

3.4. Training Data and Evaluation Metrics

The Diffusion-GAN model is trained using a suitable dataset, such as the Speech Command Zero [10] through Nine dataset, which contains digit pronunciations collected from numerous speakers. The dataset provides a diverse range of audio samples for training the model.

Method	FID ↓	IS ↑	MOS
DiffWave	169.3694	1.9737	0.29 ± 0.09
Diff-GAN (Ours)	141.8884	1.9797	3.63 ± 0.18

Table 1. Evaluation Results for the DiffWave and the Ours.

To evaluate the quality of the generated audio samples, we employed a range of evaluation metrics, including both quantitative and subjective measures.

Quantitative Evaluation We utilized commonly used metrics in generative models, including the Inception Score (IS) and the Frechet Inception Distance (FID) [3]. The IS measures the quality and diversity of the generated audio based on the predictions of a pre-trained Inception network. The FID assesses the similarity between the distributions of real and generated audio samples using the activations of the same network. Higher IS scores and lower FID scores indicate better quality and similarity to real audio, respectively.

Subjective Evaluation To account for human perception, we incorporated the Mean Opinion Score (MOS) as a subjective evaluation metric. 13 Human evaluators(but not experts) listened to the generated audio samples and provided ratings based on the overall quality, naturalness, and similarity to real audio. MOS ratings capture the subjective aspects of audio quality and offer insights that quantitative metrics may not fully capture.

By employing a combination of quantitative metrics (IS and FID) and subjective evaluation (MOS), we obtain a comprehensive understanding of the quality and perceptual experience of the generated audio samples.

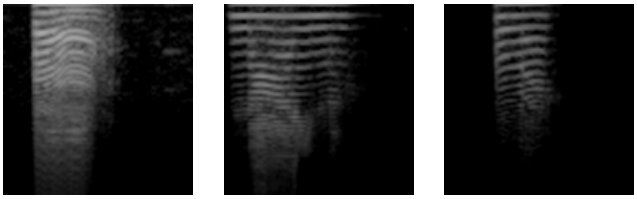
4. Experimental Results

We utilized the publicly accessible source code of DiffWave and Diffusion-GAN for our research[1] [2]. The training of the DiffWave and Diffusion GAN models was performed using the SC09 Dataset. We trained two models using an identical sample size and subsequently generated 100 audio samples. Our implementation and a subset of generated audio samples can be found on our webpage[3] for further evaluation and assessment. All experimental procedures were conducted on the GeForce RTX 3090 graphics processing unit.

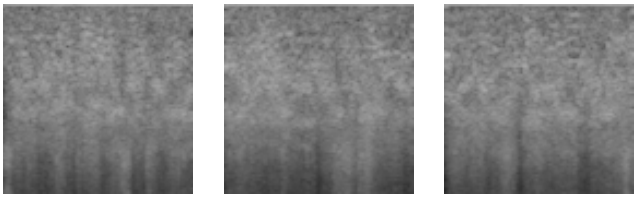
4.1. DiffWave

The unconditional version of the DiffWave model underwent a comprehensive training process consisting of 309 epochs, equivalent to 751,179 samples, over a train-

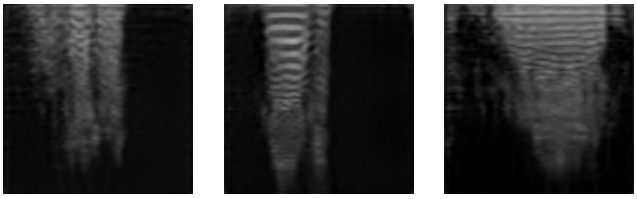
[1] https://github.com/lmnt-com/DiffWave
[2] https://github.com/Zhendong-Wang/Diffusion-GAN
[3] https://github.com/marunero/Diffusion-GAN-for-Audio-Synthesis



(a) Original



(b) DiffWave



(c) Diffusion GAN

Figure 2. Mel Spectrogram

ing period of 56 hours and 49 minutes.However, the mel-spectrogram generated by this model, as depicted in Figure 2b, fails to resemble the original dataset. It is characterized by blurriness, lacking distinct black regions and horizontal wavy stripes. Additionally, the resulting audio samples produced by the model exhibit a lack of discernible syllables and are primarily composed of noise. A human study conducted to evaluate these samples yielded an average Mean Opinion Score (MOS) of almost zero (0.29) 1, indicating an overall poor quality of the generated sound.

4.2. Diffusion-GAN

The Diffusion-GAN model was trained on a dataset with a total length of 750,000 training samples, which is comparable to the training of the DiffWave model. The training duration for Diffusion-GAN was 6 hours and 11 minutes, which is only 0.11 times the training time required by DiffWave. In Figure 2c, the generated mel-spectrogram by the Diffusion GAN model closely resembles the mel-spectrogram of the original dataset. It exhibits distinct black regions and similar horizontal wavy stripes, indicating a successful generation process. Upon examining the converted audio samples, one can readily identify discernible syllables and even recognize some words.

4.3. Comparison to DiffWave

In Table 1, we present a comprehensive comparison of the qualitative and quantitative evaluation results of the audio samples generated by DiffWave and Diffusion-GAN. Our experiment showcases the superior performance of Diffusion-GAN over the DiffWave model in terms of FID, Inception score, and MOS.

Quantitative evaluation revealed that the audio samples produced by Diffusion-GAN exhibited a lower Fréchet Inception Distance (FID) and a slightly higher Inception Score (IS) when compared to the samples generated using DiffWave. These results indicate that Diffusion-GAN generated audio samples that closely matched the distribution of real audio, exhibited higher quality and diversity.

Additionally, in a human study conducted for evaluation, the Diffusion-GAN model achieved an average Mean Opinion Score (MOS) of 3.63, indicating that it was perceived as more realistic and of higher quality by human listeners compared to the DiffWave model.

These findings demonstrate the potential of Diffusion-GAN as a powerful generative model for audio synthesis tasks. Its ability to capture the underlying distribution and produce diverse and high-quality audio samples holds promise for various applications. However, further research is needed to explore the strengths and limitations of Diffusion-GAN in different audio synthesis scenarios and to gain deeper insights into its underlying mechanisms.

5. Conclusion

In this paper, we introduced the Diffusion-GAN model for audio synthesis, which combines the principles of diffusion models and GANs to generate high-quality audio samples. Through our experiments and evaluation, we have demonstrated the effectiveness and superiority of the Diffusion-GAN model over existing approaches. Our results show that the proposed model surpasses the performance of the DiffWave model, validating its effectiveness in audio synthesis tasks.

Moving forward, further exploration and refinement of the Diffusion-GAN model are recommended. This includes investigating better evaluation criteria, exploring alternative audio embedding methods, considering different base models, and optimizing model parameters. Addressing these areas of improvement will enhance the performance and applicability of the Diffusion-GAN model in various audio synthesis tasks.

Overall, the Diffusion-GAN model holds great promise for the field of audio synthesis, and we anticipate its continued development and utilization in future studies.

References

[1] Heiga Zen Karen Simonyan Oriol Vinyals Alex Graves Nal Kalchbrenner Andrew Senior Koray Kavukcuoglu Aaron van den Oord, Sander Dieleman. Wavenet: A generative model for raw audio. 2016. 1

[2] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017. 1

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1

[5] Joo Young Hong, Jianjun He, Bhan Lam, Rishabh Gupta, and Woon-Seng Gan. Spatial audio for soundscape design: Recording and reproduction. *Applied Sciences*, 7(6), 2017. 1

[6] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. 1, 2

[7] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. 1, 2

[8] David Moffat and Mark B. Sandler. Approaches in intelligent music production. *Arts*, 8(4), 2019. 1

[9] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion, 2022. 1

[10] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209, 2018. 1, 3