SPECDIFF-GAN: A SPECTRALLY-SHAPED NOISE DIFFUSION GAN FOR SPEECH AND MUSIC SYNTHESIS

Baoueb, Teysir; Liu, Haocheng; Fontaine, Mathieu; Le Roux, Jonathan; Richard, Gaël

Abstract

Generative adversarial network (GAN) models can synthesize high- quality audio signals while ensuring fast sample generation. How- ever, they are difficult to train and are prone to several issues including mode collapse and divergence. In this paper, we introduce SpecDiff-GAN, a neural vocoder based on HiFi-GAN, which was initially devised for speech synthesis from mel spectrogram. In our model, the training stability is enhanced by means of a forward diffusion process which consists in injecting noise from a Gaussian distribution to both real and fake samples before inputting them to the discriminator. We further improve the model by exploiting a spectrally-shaped noise distribution with the aim to make the discriminator's task more challenging. We then show the merits of our proposed model for speech and music synthesis on several datasets. Our experiments confirm that our model compares favorably in audio quality and efficiency compared to several baselines.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024

SPECDIFF-GAN: A SPECTRALLY-SHAPED NOISE DIFFUSION GAN FOR SPEECH AND MUSIC SYNTHESIS

Teysir Baoueb[1], Haocheng Liu[1], Mathieu Fontaine[1], Jonathan Le Roux[2], Gaël Richard[1]

[1]LTCI, Télécom Paris, IP-Paris, France
[2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Generative adversarial network (GAN) models can synthesize high-quality audio signals while ensuring fast sample generation. However, they are difficult to train and are prone to several issues including mode collapse and divergence. In this paper, we introduce SpecDiff-GAN, a neural vocoder based on HiFi-GAN, which was initially devised for speech synthesis from mel spectrogram. In our model, the training stability is enhanced by means of a forward diffusion process which consists in injecting noise from a Gaussian distribution to both real and fake samples before inputting them to the discriminator. We further improve the model by exploiting a spectrally-shaped noise distribution with the aim to make the discriminator's task more challenging. We then show the merits of our proposed model for speech and music synthesis on several datasets. Our experiments confirm that our model compares favorably in audio quality and efficiency compared to several baselines.

Index Terms— Generative adversarial network (GAN), diffusion process, deep audio synthesis, spectral envelope

1. INTRODUCTION

Deep audio synthesis refers to a class of models which leverage neural networks to generate natural-sounding audio signals based on given acoustic features. It has applications in many different tasks including the generation of speech (e.g., text-to-speech (TTS) [1], speech-to-speech translation [2], voice conversion [3]), music synthesis [4, 5], and sound effects generation [6, 7].

Audio synthesis was for long dominated by likelihood-based models such as autoregressive models [8] and flow-based models [9]. However, the sequential nature of the former models leads to slow inference times as each output element is generated one by one, conditioned on previously generated elements. Flow-based models, on the other hand, are not parameter-efficient as they typically require a deep architecture to perform complex invertible transformations.

With the emergence of generative adversarial networks (GANs) [10], which have yielded promising results in the generation of high-resolution images, GAN-based audio synthesis models have been proposed [11, 12]. They can produce high-fidelity waveforms while maintaining a fast and computationally competitive sampling. However, GANs are hard to train and are known to suffer from mode collapse [13]. This issue was addressed by denoising diffusion probabilistic models (DDPMs) [14–16], but these models suffer themselves from a slow reverse process, which requires a huge number of
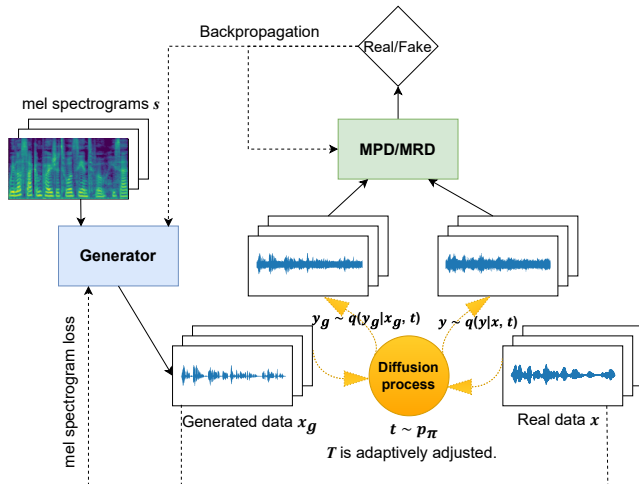
Fig. 1: Overview of SpecDiff-GAN

steps to obtain satisfactory results, thus making them inapplicable in real-life settings.

In this paper, we propose to tackle the training instability of GANs and the slow inference process of DDPMs. To that aim, we choose HiFi-GAN [11], an efficient and high-quality mel spectrogram to speech waveform synthesizer, as a core model, and build an enhanced HiFi-GAN model exploiting a noise-shaping diffusion process, showing the merit of our proposed model on a large variety of audio signals. More precisely, our main contributions include:

- The injection of instance noise into both inputs (real and fake) of the discriminator similarly to [17] to help stabilize the training;

- The use of a spectrally-shaped noise distribution to make the discriminator's task more challenging. In particular, we evaluate several variations for the noise distribution exploiting the inverse filter described in [18], which is based on the spectral envelope of the mel spectrogram input;

- An extensive experimental work with application not only to speech but also instrumental music synthesis, which to the best of our knowledge has not been done before for Hifi-GAN based models.

Our proposed model is illustrated in Fig. 1. Examples and full code are available at `https://specdiff-gan.github.io/`.

2. RELATED WORK

2.1. HiFi-GAN

HiFi-GAN [11] addresses the challenges of high-quality speech synthesis by leveraging GANs. The model employs a generator network

that takes mel spectrograms as input and utilizes a progressive up-sampling process to synthesize time-domain waveforms closely resembling the original audio signals. HiFi-GAN's architecture features a multi-receptive field fusion module, which enhances representation by integrating information from different receptive regions. Additionally, it features two discriminators: multi-period discriminator (MPD) and multi-scale discriminator (MSD), which respectively capture periodic patterns and identify long-term dependencies. This approach has demonstrated remarkable performance in generating high-quality audio with improved sampling accuracy and speed.

2.2. SpecGrad

SpecGrad, introduced by Koizumi et al. [18], is a diffusion-based vocoder. This model enhances the quality of synthesized audio by leveraging a diffusion process that adapts the shaping of noise in the spectral domain. Let $\mathcal{N}(0, \boldsymbol{\Sigma})$ be the noise distribution. SpecGrad proposes to include information from the spectral envelope into $\boldsymbol{\Sigma}$. To achieve this, $\boldsymbol{\Sigma}$ is decomposed as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L} = \boldsymbol{G}^+ \boldsymbol{M}_{\mathrm{SG}} \boldsymbol{G}$, with $\boldsymbol{G}$ and $\boldsymbol{G}^+$ denoting matrix representations of the short-time Fourier transform (STFT) and its inverse, and $\boldsymbol{M}_{\mathrm{SG}}$ a complex diagonal matrix representing a filter based on the spectral envelope. Specifically, the magnitude of $\boldsymbol{M}_{\mathrm{SG}}$ aligns with the spectral envelope, while the phase component is obtained as that of the minimum phase response. By incorporating spectral envelope information in this way, SpecGrad enhances the modeling of audio signals, resulting in improved audio quality and naturalness in the generated audio compared to previous diffusion models [15, 16]. However, it is important to note that the slow inference speed of SpecGrad limits its suitability for real-world applications.

2.3. Diffusion-GAN

Diffusion-GAN [17] is a novel approach for training GANs using diffusion processes to enhance stability and quality. By gradually transforming real and generated samples through an adaptive diffusion process, Diffusion-GAN bridges the gap between initial generator outputs and the target data distribution. This regularization mechanism mitigates challenges associated with mode collapse and unstable training dynamics, contributing to improved training efficiency and sample quality in GANs. The original paper applied this approach to image synthesis, and its application to the audio domain remains limited [19].

3. PROPOSED METHOD

3.1. Architecture

Our generator network closely mirrors the architecture used in HiFi-GAN, chosen for its remarkable capability to produce high-quality audio samples swiftly. Furthermore, we incorporate HiFi-GAN's multi-period discriminator (MPD), which comprises several sub-discriminators, each parameterized with a period p, to effectively capture periodic patterns. However, instead of utilizing the multi-scale discriminator (MSD), we opted for UnivNet's multi-resolution discriminator (MRD) [20]. MRD is a composition of multiple sub-discriminators, each parameterized by a tuple indicating (FFT size, hop size, Hann window length). These varying temporal and spectral resolutions enable the generation of high-resolution signals across the full band. Integrating MRD consistently improves sample quality and reduces artefacts in audio synthesis, as shown in [12,21].

3.2. Enhancing the GAN model with diffusion

Following [17], we leverage a diffusion process during GAN training. In this approach, rather than discerning between the original and generated data, the discriminator learns to distinguish between the perturbed versions of each (see Fig. 1).

We recall that, during the forward diffusion process, an initial sample denoted as $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ undergoes a series of T sequential steps where it is progressively perturbed by Gaussian noise. Denoting the noise schedule by $\{\beta_t\}_{t=1}^{T}$, this can be formalized as $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t \geq 1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ with $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I})$. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{u=1}^{t} \alpha_u$. It can be shown that, in the forward process, $\boldsymbol{x}_t$ can be sampled at any arbitrary time step t in closed form by $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

In the context of our model, we denote by $\boldsymbol{x} \sim p(\boldsymbol{x})$ the ground-truth audio and by $\boldsymbol{s}$ the input condition of the generator, i.e., the mel spectrogram of the ground-truth audio. Using these notations, $G(\boldsymbol{s})$ is the generated signal. Perturbed samples are acquired as follows:

$$\boldsymbol{y} \sim q(\boldsymbol{y}|\boldsymbol{x}, t), \quad \boldsymbol{y} = \sqrt{\bar{\alpha}_t}\boldsymbol{x} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \qquad (1)$$

$$\boldsymbol{y}_g \sim q(\boldsymbol{y}_g|G(\boldsymbol{s}), t), \quad \boldsymbol{y}_g = \sqrt{\bar{\alpha}_t}G(\boldsymbol{s}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}' \qquad (2)$$

where $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $q(\boldsymbol{y}|\boldsymbol{x}, t)$ is the conditional distribution of the noisy sample $\boldsymbol{y}$ given the target data $\boldsymbol{x}$ and the diffusion step t and $q(\boldsymbol{y}_g|G(\boldsymbol{s}), t)$ is the conditional distribution of the noisy sample $\boldsymbol{y}_g$ given the generated signal $G(\boldsymbol{s})$ and the diffusion step t.

3.3. Noise distribution

We explore two options for $\boldsymbol{\Sigma}$. In the first case, we set it to $\boldsymbol{\Sigma}_{\mathrm{standard}} = \sigma^2 \mathbf{I}$, a similar approach to that in [17], where I represents the identity matrix and σ is a scalar. We refer to this model as StandardDiff-GAN. In the second option, drawing inspiration from SpecGrad [18], we shape the noise based on the spectral envelope. Our filter $\boldsymbol{M}_{\mathrm{spec}}$ is however the inverse of the one used in SpecGrad, specifically $\boldsymbol{M}_{\mathrm{spec}} = \boldsymbol{M}_{\mathrm{SG}}^{-1}$. This choice results in a noise distribution that emphasizes increased noise incorporation in low-energy regions, thereby challenging the discriminator. The version of our model incorporating this noise distribution, with variance $\boldsymbol{\Sigma}_{\mathrm{spec}} = \boldsymbol{L}_{\mathrm{spec}}\boldsymbol{L}_{\mathrm{spec}}^T$ where $\boldsymbol{L}_{\mathrm{spec}} = \boldsymbol{G}^+ \boldsymbol{M}_{\mathrm{spec}} \boldsymbol{G}$, is referred to as SpecDiff-GAN.

3.4. Adaptive diffusion

Similar to the approach in [17], we dynamically regulate the level of difficulty for the discriminators during training by incorporating an adaptive update mechanism for the maximum number of diffusion steps, denoted as T, within the interval $[T_{\min}, T_{\max}]$. This adaptive adjustment ensures that the discriminators are provided with varying degrees of challenge as they learn to distinguish between real and generated samples. When the discriminators struggle to perform effectively, we decrease T to provide more opportunities for learning from relatively simpler samples, such as non-perturbed or slightly noisy ones. Conversely, if the discriminators find it too easy to differentiate between the diffused generated and real samples, we increase T to introduce more complexity to their task.

To quantify the extent of discriminator overfitting to the training data, we employ a metric similar to that in [22], computed over B consecutive minibatches as

$$r_d = \mathbb{E}[\mathrm{sign}(D_{\mathrm{train}} - 0.5)], \qquad (3)$$

where D_{train} represents the discriminator outputs on samples of the training set and $\mathbb{E}[\cdot]$ a mean over the B minibatches. r_d attempts to estimate the portion of the training set for which discriminator outputs would exceed 0.5. A value of r_d close to 1 indicates overfitting, while a value close to 0 suggests no overfitting. We update T every $B = 4$ minibatches using the following rule:

$$T \leftarrow T + \text{sign}(r_d - d_{\text{target}}) \cdot C, \tag{4}$$

where d_{target} is a hyperparameter representing the desired value for r_d, and C is a constant chosen to regulate the rate at which T transitions from $T_{\min}$ to $T_{\max}$. The diffusion timestep $t \leq T$ is then drawn from a discrete distribution p_π defined with $c_T = \sum_{u=1}^{T} u$ as:

$$t \sim p_\pi := \text{Discrete}\left(1/c_T, 2/c_T, \ldots, T/c_T\right). \tag{5}$$

This distribution gives more weight to larger values of t, influencing the choice of diffusion steps during training.

3.5. Training losses

We here describe the various training losses. For the sake of simplicity, we denote both discriminators as D following [11].

Our discriminative loss is provided by the following formula:

$$\mathcal{L}_D = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s},t,\boldsymbol{y},\boldsymbol{y}_g)}\left[(D(\boldsymbol{y}) - 1)^2 + (D(\boldsymbol{y}_g))^2\right], \tag{6}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}_g$ are obtained as in Section 3.2. For HiFi-GAN, the loss is simply obtained as $\mathcal{L}_D = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}[(D(\boldsymbol{x}) - 1)^2 + (D(G(\boldsymbol{s})))^2]$.

The SpecDiff-GAN generator, as HiFi-GAN, employs an adversarial loss and two extra losses to enhance perceptual and spectral similarity with the ground-truth audio, a feature matching (FM) loss and a mel spectrogram loss. The total loss is formulated as

$$\begin{aligned} \mathcal{L}_G = {} & \mathbb{E}_{(\boldsymbol{s},t,\boldsymbol{y}_g)}\left[(D(\boldsymbol{y}_g) - 1)^2\right] \\ & + \lambda_{\text{FM}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s},t,\boldsymbol{y},\boldsymbol{s},\boldsymbol{y}_g)}\left[\sum_{i=1}^{L} \frac{1}{N_i} \|D^i(\boldsymbol{y}) - D^i(\boldsymbol{y}_g)\|_1\right] \\ & + \lambda_{\text{mel}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})}\left[\|\phi(\boldsymbol{x}) - \phi(G(\boldsymbol{s}))\|_1\right], \tag{7} \end{aligned}$$

where λ_{FM} and λ_{mel} are scalar coefficients, ϕ is a function that transforms a waveform into its mel spectrogram, L denotes the number of layers in the discriminator, D^i the features in the ith layer of the discriminator and N_i their number. It is important to highlight that we employ the diffused versions of the real and fake data only for the first two terms in the generator loss. The last term in Eq. (7) indeed does not involve the discriminator.

4. EXPERIMENTS

We present hereafter the experimental protocol used to evaluate our method and the baseline models.

4.1. Datasets

For our experiments, we consider the following datasets:
- **LJSpeech** [23] is a single-speaker speech dataset. It contains English recordings sampled at 22050 Hz with a total duration of ~ 24 hours. We use the same train/test split as in HiFi-GAN [11] (i.e., 12950 clips for training and 150 clips for testing).

- **VCTK** [24] is a clean multispeaker dataset with 110 speakers, 63 female and 47 male. The clips were recorded using two microphones and we consider the Microphone 1 configuration. It comprises ~ 41 hours of utterances in different English accents. We resample the recordings from 48 kHz to 24 kHz. We keep 10 speakers for testing and use the others for training.
- **MAPS** [25] is a dataset of MIDI piano recordings captured under 9 distinct recording conditions, and sampled at a rate of 44.1 kHz. We focus on a specific subset consisting of classical piano compositions (MUS), totaling approximately 18 hours. We split the dataset into 229 pieces for training and 41 pieces for testing. Subsequently, we converted offline all tracks to single-channel audio and segmented them into 5-second fragments.
- **ENST-Drums** [26] contains recordings by 3 drummers on 8 individual audio channels with a total duration of 225 minutes. The tracks were recorded using various drum kits and are sampled at 44.1 kHz. We split the recordings into 2512 for training and 466 for testing. Similarly to MAPS, our pre-processing pipeline involves an offline conversion from stereo to mono and the subsequent segmentation of audio clips.

4.2. Model Setup

For MRD, we incorporate 3 sub-discriminators with the same parameters as [20]: $(1024, 120, 600)$, $(2048, 240, 1200)$, and $(512, 50, 240)$. As in [11], we consider 5 sub-discriminators for MPD with periods 2, 3, 5, 7, and 11 to prevent overlaps, and $\lambda_{\text{FM}} = 2$ and $\lambda_{\text{Mel}} = 45$ for the generator loss. We also keep the same choice of optimizer and learning rate scheduler. For the diffusion process, we adopt $d_{\text{target}} = 0.6$, where experiments with other values showed no significant difference, and $\sigma = 0.05$ as per [17].

We compare our model to other GAN models, namely HiFi-GAN[1], UnivNet-c32[2] and BigVGAN base model[3].

4.3. Training configurations

Detailed training configurations for each dataset across the various models are as follows:
- **LJSpeech:** The parameter values chosen are consistent with the V1 configuration of HiFi-GAN. The initial learning rate is set to $2 \cdot 10^{-4}$ across all models, except for UnivNet and BigVGAN, where we conduct experiments with an initial learning rate of 10^{-4} in accordance with the settings outlined in their respective papers.
- **VCTK:** We adopt the 24 kHz base configuration of BigVGAN. The other parameters are the same as those used for LJSpeech.
- **MAPS** and **ENST-Drums**: We use 128-dimensional log-mel spectrograms with a Hann window size of 2048, a frame shift of 512, and 2048-point FFT with a full-band range (0 - 22.050 kHz). UnivNet is not used in this configuration due to code adjustments required. For other models, we increase upsampling rates and kernel sizes to $[8, 8, 2, 2, 2]$ and $[16, 16, 4, 4, 4]$ respectively. All models are trained with an initial learning rate of $2 \cdot 10^{-4}$ and a segment size of 16384.

All models are trained on 1 NVIDIA A100 GPU for 1M steps, with a batch size of 16. All generators have approximately 14M parameters.

---

[1] https://github.com/jik876/hifi-gan
[2] https://github.com/maum-ai/univnet
[3] https://github.com/NVIDIA/BigVGAN

Table 1: Inference results on LJSpeech. (lr: initial learning rate)

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
HiFi-GAN	3.47	0.98	1.203
UnivNet (lr=1e-4)	3.44	0.98	1.330
StandardDiff-GAN	3.62	0.98	1.086
SpecDiff-GAN	**3.76**	**0.99**	**1.018**
BigVGAN (lr=1e-4)	3.72	0.98	1.073

Table 2: Inference results on VCTK. (lr: initial learning rate)

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
HiFi-GAN	2.97	0.94	1.213
UnivNet (lr=1e-4)	3.21	0.94	1.209
StandardDiff-GAN	3.37	0.96	1.046
SpecDiff-GAN	3.52	0.96	0.983
BigVGAN (lr=1e-4)	3.67	0.96	0.959

Table 3: FAD (↓) scores on MAPS and ENST-Drums datasets.

Model	MAPS	ENST-Drums
HiFi-GAN	0.153	0.226
StandardDiff-GAN	0.108	**0.138**
SpecDiff-GAN	0.080	0.149
BigVGAN	**0.075**	0.190

5. RESULTS

To evaluate the performance of trained models, we use Perceptual Evaluation of Speech Quality (PESQ) [27], Short-Time Objective Intelligibility (STOI) [28] and WARP-Q [29] for speech synthesis. For each metric, we report the mean of the scores over all the pieces in the test set. Each 95% confidence interval around the mean value has margins smaller than 0.03, 0.001, and 0.008 respectively. For music generation, we utilize the Fréchet Audio Distance (FAD) [30] with the VGGish model [31] to generate the embeddings.

5.1. Inference results for the different datasets

Table 1 presents the results on LJSpeech. SpecDiff-GAN exhibits superior performance in terms of audio quality and speech intelligibility when compared to both the baseline models and BigVGAN. Performance drops with $\Sigma_{standard}$ (StandardDiff-GAN), highlighting the importance of the noise shaping. Our model excels with known speakers in band-limited conditions during inference.

The results for the VCTK dataset, which involves inference on unseen speakers, are reported in Table 2. Among the models, BigV-GAN with a learning rate of $2 \cdot 10^{-4}$ has the best performance. SpecDiff-GAN closely follows, with a negligible difference that is not statistically significant. It is noteworthy that both SpecDiff-GAN and StandardDiff-GAN outperform the baseline models, HiFi-GAN and UnivNet. In particular, SpecDiff-GAN showcases a substantial performance margin compared to the baselines.

Table 3 displays the results for the MAPS and ENST-Drums datasets. For MAPS, SpecDiff-GAN outperforms both HiFi-GAN and StandardDiff-GAN, highlighting the advantage of employing the spectrally-shaped noise distribution. Notwithstanding, BigVGAN demonstrates a slightly better performance compared to SpecDiff-GAN. Surprisingly, in the case of ENST-Drums, StandardDiff-GAN outperforms the other models, with SpecDiff-GAN following closely behind. The ENST-Drums dataset's small size (225 minutes) and multiple tracks for the same drum performance from different channels may have hindered the learning process.

Table 4: Ablation study on LJSpeech.

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
SpecDiff-GAN	**3.76**	**0.99**	**1.018**
StandardDiff-GAN	3.62	0.98	1.086
Without diffusion	3.52	0.98	1.135
MRD → MSD	3.65	0.98	1.069
$(\Sigma_{spec} \to \Sigma_{standard})$ + (MRD → MSD)	3.54	0.98	1.156

Table 5: Synthesis speed compared to real-time evaluated with a batch of 100 one-second-long samples on 1 NVIDIA V100 GPU

Model	LJSpeech	VCTK	MAPS	ENST-Drums
BigVGAN base	× 23.28	× 21.40	× 18.03	× 18.03
SpecDiff-GAN	× 220.96	× 203.28	× 183.46	× 183.15

5.2. Ablation study

We conducted an ablation study on the MRD, the diffusion process, and the reshaped noise distribution to assess the individual impact of each component on the quality of the generated audio. We train all models on LJSpeech for 1M steps. The results are presented in Table 4. Eliminating the spectrally-shaped noise distribution and adopting $\Sigma_{standard}$ instead (StandardDiff-GAN) leads to a deterioration in results. This behaviour is also observed when replacing the MRD with the MSD. Furthermore, when we exclude the diffusion process entirely ("Without diffusion") or maintain it with $\Sigma_{standard}$ and substitute the MSD for the MRD, the results decline even further. It is worth noting that the last row in the table is equivalent to HiFi-GAN with a non-spectrally-shaped diffusion process. A comparison of metric scores with those of HiFi-GAN in Table 1 reveals that the diffusion process leads to improvement. This highlights that each component of our model (MRD, diffusion, shaped noise) plays a crucial role in enhancing audio quality.

5.3. Model complexity

In comparison to the base BigVGAN model, our model features approximately 200k fewer parameters for all tested configurations. Furthermore, our model demonstrates a notably faster synthesis speed, as detailed in Table 5. This speed is equivalent to that of HiFi-GAN and StandardDiff-GAN since they all share the same generator. The primary reason for BigVGAN's slower performance lies in its utilization of the computationally intensive snake activation function [32]. This characteristic also makes BigVGAN significantly slower to train compared to our model, with a training duration factor ranging from 1.5 to 2.

6. CONCLUSION

We introduced SpecDiff-GAN, a novel approach harnessing a forward diffusion process with spectrally-shaped noise to enhance GAN-based audio synthesis. Our application spanned both speech and music generation. The experimental results showcased SpecDiff-GAN's capacity to generate high-quality waveforms surpassing baselines while being competitive to the state-of-the-art model, BigVGAN. Notably, SpecDiff-GAN maintained efficient inference speeds. Our approach is versatile, offering adaptability to various GAN-based audio synthesis models.

Future research avenues include testing our model on a larger, more diverse dataset, covering a wide spectrum of sound types for universal audio synthesis.

7. REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.

[2] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen *et al.*, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. Interspeech*, 2019.

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 132–157, 2020.

[4] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," *arXiv preprint arXiv:1902.08710*, 2019.

[5] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801*, 2020.

[6] D. Moffat, R. Selfridge, and J. Reiss, "Sound effect synthesis," in *Foundations in Sound Design for Interactive Media*, M. Filimowicz, Ed. Routledge, 2019.

[7] C. Schreck, D. Rohmer, D. L. James, S. Hahmann, and M.-P. Cani, "Real-time sound synthesis for paper material based on geometric analysis," in *Proc. ACM SIGGRAPH/Eurographics SCA*, Jul. 2016.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves *et al.*, "WaveNet: A generative model for raw audio," in *Proc. ISCA SSW*, 2016.

[9] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, May 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[12] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. ICLR*, 2023.

[13] N. Kodali, J. Hays, J. Abernethy, and Z. Kira, "On convergence and stability of GANs," *arXiv preprint arXiv:1705.07215*, 2017.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[15] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, 2021.

[16] S. gil Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng *et al.*, "PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *Proc. ICLR*, 2022.

[17] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with diffusion," in *Proc. ICLR*, 2023.

[18] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *Proc. Interspeech*, 2022.

[19] X. Wu, "Enhancing unsupervised speech recognition with diffusion GANs," in *Proc. ICASSP*, Jun. 2023.

[20] W. Jang, D. C. Y. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proc. Interspeech*, 2021.

[21] C. Wang, C. Zeng, and X. He, "HiFi-WaveGAN: Generative adversarial network with auxiliary spectrogram-phase loss for high-fidelity singing voice generation," *arXiv preprint arXiv:2210.12740*, 2022.

[22] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. NeurIPS*, 2020.

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

[24] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR), Tech. Rep., 2019.

[25] V. Emiya, N. Bertin, B. David, and R. Badeau, "MAPS - a piano database for multipitch estimation and automatic transcription of music," INRIA, Tech. Rep., Jul. 2010.

[26] O. Gillet and G. Richard, "ENST-Drums: An extensive audio-visual database for drum signals processing," in *Proc. ISMIR*, 2006.

[27] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[29] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Warp-Q: Quality prediction for generative neural speech codecs," *Proc. ICASSP*, pp. 401–405, 2021.

[30] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019.

[31] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, 2017.

[32] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," in *Proc. NeurIPS*, 2020.