FASTDIFF 2: DUALLY INCORPORATING GANS INTO DIFFUSION MODELS FOR HIGH-QUALITY SPEECH SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

FastDiff, as a class of denoising probabilistic models, has recently achieved impressive performances in speech synthesis. It utilizes a noise predictor to learn a tight inference schedule for skipping denoising steps. Despite the successful speedup of FastDiff, there is still room for improvements, e.g., further optimizing the speed-quality trade-off and accelerating DDPMs training procedures. After analyzing GANs and diffusion models in conditional speech synthesis, we find that: GANs produce samples but do not cover the whole distribution, and the coverage degree does not distinctly impact audio quality. Inspired by these observations, we propose to trade off diversity for quality and speed by incorporating GANs into diffusion models, introducing two GAN-empowered modeling perspectives: (1) FastDiff 2 (Diff-GAN), whose denoising distribution is parametrized by conditional GANs; and (2) FastDiff 2 (GAN-Diff), in which the denoising model is treated as a generator in GAN for adversarial training. Unlike the acceleration methods based on skipping the denoising steps, FastDiff 2 provides a principled way to speed up both the training and inference processes. Experimental results demonstrate that both variants of FastDiff 2 enjoy an efficient 4-step sampling process as in FastDiff yet demonstrate a superior sample quality.[1]

1 INTRODUCTION

With the recent development of deep generative models, speech synthesis has seen extraordinary progress. Previous models (Oord et al., 2016; Kalchbrenner et al., 2018) generate waveforms autoregressively from mel-spectrograms yet suffer from slow inference speed. Non-autoregressive methods (Prenger et al., 2019; Kumar et al., 2019; Kong et al., 2020b) such as Generative adversarial network (GAN) (Creswell et al., 2018; Mao et al., 2019) have been designed to address this issue, they generate samples with extremely fast speed and achieve comparable voice quality with autoregressive models.

Recently, a class of generative models called denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Song et al., 2020) have emerged and demonstrated surprisingly good results in both image and audio synthesis. However, a guarantee of high sample quality typically comes at the cost of thousands of diffusion iterations, making their application expensive in practice. Among speech diffusion models (Liu et al., 2021; Popov et al., 2021; Chen et al., 2020), FastDiff (Huang et al., 2022) conducts extensive investigations on accelerating the sampling process and utilizes a noise predictor (Lam et al., 2022) to learn a tight inference schedule for skipping denoising steps. This design demonstrates the impressive results and makes diffusion models more applicable to real-world deployment at a low computational cost, while there is still room for improvements: 1) further optimizing the speed-quality trade-off; and 2) accelerating training procedures for estimating data density gradient.

Through some preliminary experiments and analyses (see Section 3), we observe that: 1) GANs tend to generate high-quality samples but do not cover the whole distribution, which sacrifices sample diversity for quality and speed; and 2) distribution coverage degree does not distinctly impact

[1] Audio samples are available at `https://FastDiff2.github.io/`

audio quality. Inspired by this, we propose to trade off diversity for better quality and speed by incorporating GANs into diffusion models, introducing two GAN-empowered modeling perspectives: 1) FastDiff 2 (Diff-GAN): a diffusion model whose denoising process is parametrized by conditional GANs, and the non-Gaussian denoising distribution makes it much more stable to implement the reverse process with large steps sizes; and 2) FastDiff 2 (GAN-Diff): a generative adversarial network whose forward process is constructed by multiple denoising diffusion iterations, which exhibits better training stability and mode coverage.

Unlike the acceleration strategies based on jumping denoising steps that only accelerate the generation process during inference, FastDiff 2 provides a principled way to accelerate DDPMs in training and inference. Experimental results show that both variants of FastDiff 2 enjoy an effective 4-iter sampling process as in FastDiff yet demonstrate the outperformed sample quality. We further show that FastDiff 2 generalizes well to the mel-spectrogram inversion of unseen speakers. The main contributions of this work are summarized as follows:

- We investigate two popular classes of deep generative models (diffusion models and GANs) in conditional speech synthesis, and observe that GANs produce samples but do not cover the whole distribution, and coverage degree does not distinctly impact audio quality. Inspired by these, we propose to trade off diversity for quality and speed by incorporating GANs into diffusion models, proposing FastDiff 2 with two GAN-empowered modeling perspectives.
- FastDiff 2 (Diff-GAN) removes the common assumption of Gaussian distribution and utilizes conditional GANs to parametrize the multimodal denoising distribution, allowing to implement the reverse process with large step sizes more stably.
- FastDiff 2 (GAN-Diff) breaks the one-shot forward of conditional GANs into several denoising diffusion steps in which each step is relatively simple to model, and thus it exhibits better training stability and mode coverage.
- Experimental results show that both variants of FastDiff 2 enjoy an effective 4-iter sampling process as in FastDiff yet demonstrate a superior sample quality, providing a principled way to accelerate DDPMs in training and inference.

2 Background on Speech Synthesis

With the development of deep generative models, speech synthesis technology has made rapid progress up to date. Most models (Wang et al., 2017; Ren et al., 2019) first convert input text or phoneme sequence into mel-spectrogram, and then transform it to waveform using a separately trained vocoder (Kumar et al., 2019; Kong et al., 2020a). In this work, we focus on designing the second-stage model that efficiently synthesizes high-fidelity waveforms from mel-spectrograms.

Neural vocoders require diverse receptive field patterns to catch audio dependencies, and thus previous models (Oord et al., 2016; Kalchbrenner et al., 2018) generate waveforms autoregressively from mel-spectrograms yet suffer from slow inference speed. In recent years, non-autoregressive methods (Prenger et al., 2019; Kumar et al., 2019; Kong et al., 2020b) have been designed to address this issue, which generates samples with extremely fast speed while achieving comparable voice quality with autoregressive models. Below we mainly introduce two popular classes of deep generative models (diffusion models and GANs) for conditional speech synthesis:

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs) (Kumar et al., 2019; Kong et al., 2020a) are one of the most dominant non-autoregressive models in speech synthesis. GANs jointly train a powerful generator G and discriminator D with a min-max game:

$$\min_G \max_D V(G,D) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1-D(G(\mathbf{z})))], \tag{1}$$

The generator G aims to transform noise z into $G(z)$ that mimics real data, while the discriminator D learns to distinguish the generated samples $G(z)$ from real ones. Morrison et al. (2021) propose a chunked autoregressive GAN for conditional waveform synthesis, Lee et al. (2022) utilize a large-scale pretraining to improve out-of-distribution quality, Bak et al. (2022) investigate GAN-based neural vocoders and proposes an artifact-free GAN-based neural vocoder.

However, GAN-based models are often difficult to train, collapsing (Creswell et al., 2018) without carefully selected hyperparameters and regularizers, and showing less sample diversity.

2.2 DIFFUSION PROBABILISTIC MODELS

Recently proposed denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) are likelihood-based generative models that have recently succeeded in advancing the state-of-the-art results in most image and audio synthesis tasks. Given data distribution as $q(\mathbf{x}_0)$, the diffusion process is defined by a fixed Markov chain from data $\mathbf{x}_0$ to the latent variable $\mathbf{x}_T$: there is a forward process that gradually adds noise to the data $q(\mathbf{x}_0)$ in T steps with pre-defined noise schedule β_t:

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{2}$$

The reverse process is to recover samples from Gaussian noises parameterized by shared θ. A guarantee of high sample diversity typically comes at the cost of hundreds to thousands of denoising steps:

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{3}$$

It has been demonstrated that diffusion probabilistic models (Dhariwal & Nichol, 2021; Xiao et al., 2021) can learn diverse data distribution in multiple domains, such as images and time series. However, an apparent degradation could be witnessed when reducing reverse iterations, making them challenging to get accelerated.

3 PRELIMINARY STUDY

In image generation, superior sample diversity (Dhariwal & Nichol, 2021; Ho et al., 2020; Song et al., 2020) is a crucial reason for the diffusion model in producing high-quality samples even on the challenging dataset. Due to the distinctive advantages of diversity and distribution coverage over GANs, diffusion models have demonstrated to generate realistic and vivid images, achieving the current state-of-the-art measured by FID.

Diffusion literature mainly demonstrates the benefit of distribution coverage in generating individual images, while audio data is different (Oord et al., 2016; Kalchbrenner et al., 2018) for its long-term dependencies, high sampling rate, and strong condition. In this section, we focus on investigating the characteristic of diffusion models and GANs with close model capacity in conditional speech synthesis, which are relatively overlooked. Specifically, we evaluate the performance (including sample quality, speed, and diversity) of generated samples and explore how distribution coverage impacts sample quality by auditory sensation.

3.1 EXPERIMENTAL SETUP

We prepare 20 unseen samples from the benchmark LJSpeech dataset (Ito & Johnson, 2017) for evaluation. For a fair comparison, we implement the GAN and diffusion model with a shared backbone (Huang et al., 2022), which comprises three Diffusion-UBlock and DBlock with the up/downsample rate of [8, 8, 4]. Following the common practice (Kumar et al., 2019; Yamamoto et al., 2020), we remove the time embedding in GAN and introduce an auxiliary multi-resolution STFT loss to stabilize adversarial learning. More information has been attached in Appendix D.1.

3.2 VISUALIZATION

We further visualize the marginal distributions $P(\mathbf{x}|ph)$ of diffusion models and GANs in Figure 1. Specifically, we 1) randomly sample 100 latent noises z for each testing audio and obtain 2000 utterances in total. 2) split the generated utterances into phoneme-level samples according to the boundary obtained by forced alignment (McAuliffe et al., 2017) and transform them into linear spectrograms; 3) compute the histograms [2] and smooth them into probability density functions with kernel density estimation for better visualization.

[2] We obtain similar results among different frequency bands and choose the 70-th bin for illustration.

Table 1: Comparison of GANs and diffusion models for speech synthesis. We crowd-source 5-scale MOS tests via Amazon Mechanical Turk, which are recorded with 95% confidence intervals (CI). We implement real-time factor (RTF) assessment on a single NVIDIA V100 GPU.

Model	**Quality**			**Speed**	**Diversity**	
	MOS (↑)	MCD (↓)	PESQ (↑)	RTF (↓)	NDB (↓)	JS (↓)
GT	4.32±0.06	/	/	/	/	/
GAN	4.08±0.07	**1.48**	3.87	**0.001**	34	0.0016
Diffusion	**4.16±0.09**	1.62	**3.92**	4.70	**22**	**0.0010**

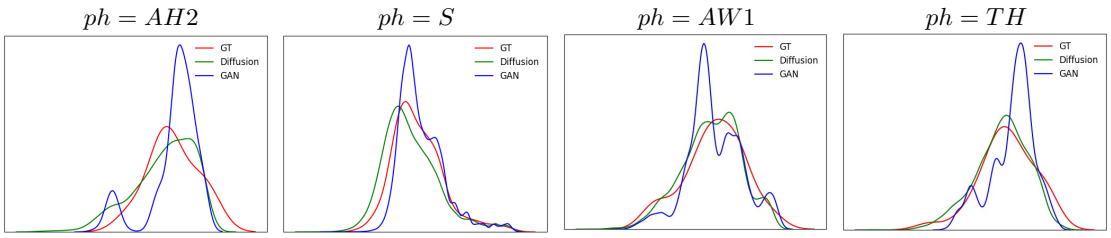


Figure 1: Comparison of sample distribution coverage between diffusion models and GANs. We randomly choose 4 different phonemes ($ph = AH2, S, AW1, TH$) in this case study.

3.3 Analyses

Based on the evaluation results presented in Table 1 and the marginal distributions illustrated in Figure 1, we have the following observations:

Diffusion models demonstrate better sample diversity. A more diverse data distribution could be observed in samples generated by diffusion models, demonstrating a better mode convergence. Diffusion models are better at data sharpness, diversity, and matching marginal label distribution of training data. However, sampling from diffusion models often requires thousands of network evaluations, which is significantly slower than GAN and makes their application expensive in practice.

GANs trade off diversity for quality and speed. A distinct degradation of mode convergence could be witnessed in GANs, which tend to produce samples but do not cover the whole distribution, indicating a collapsed distribution and less sample diversity. To conclude, GANs sacrifice diversity for quality and speed, and the constrained learned distribution does not hinder their ability to generate high-quality samples. Compared to diffusion models, GANs enjoy high-quality speech synthesis with a minor gap of 0.08 in MOS, while even demonstrating an outperformed performance in MCD evaluation. Regarding inference speed, GANs enjoy an effective one-shot sampling process, significantly reducing the inference time compared with competing diffusion mechanisms.

Distribution coverage is not equivalent to sample quality by auditory sensation. In long continuous signals, sample diversity mainly resorts to the local fluctuation and phase variation predicted given amplitude condition (intermediate acoustic feature). Different from the visual diversity in images, people hardly discriminate varied patterns in the long continuous waveforms (typically 16000 samples per second), and thus the restricted sample distribution does not necessarily mean a decreased audio quality. The high computational cost of diffusion models for increasing diversity has become a heavy burden especially when the sampling rate of audio is high.

4 FastDiff 2

After analyzing two kinds of generative NAR models (GAN and diffusion) for speech synthesis, we witness that GANs sacrifice sample diversity for better quality and speed, producing high-quality samples but not covering the whole distribution. As such, we aim to acquire this benefit (i.e., trade off diversity for quality and speed) by incorporating GANs into diffusion models.

4.1 Overview

This section presents our proposed FastDiff 2, including two GAN-empowered modeling perspectives: 1) FastDiff 2 (Diff-GAN): a diffusion model whose denoising process is parametrized by

conditional GANs, and thus the non-Gaussian denoising distribution makes it much more stable to implement the reverse process with a large step size; and 2) FastDiff 2 (GAN-Diff): a generative adversarial network whose forward process is constructed by multiple denoising diffusion distributions, thus exhibiting better training stability and mode convergence.

4.2 DIFFUSION MECHANISM LEVERAGING GAN

Diffusion models commonly assume that the denoising distribution can be approximated by Gaussian distributions. However, the Gaussian assumption holds only in the infinitesimal limit of small denoising steps, leading to the requirement of numerous steps in the reverse process. As such, reducing the number of iterative steps always causes a distinct degradation in perceptual quality.

In this work, we propose **FastDiff 2 (Diff-GAN)** leveraging conditional GANs to model the denoising distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the non-Gaussian multimodal distribution makes it much more stable to implement the reverse process with large steps sizes. Specifically, our forward diffusion process is set up with the main assumption that the number of diffusion iterations is small ($T = 4$). The training is formulated by matching the conditional GAN generator $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ using an adversarial loss that minimizes a divergence D_{adv} per denoising step. The discriminator $D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)$ is designed to be diffusion-step-dependent, which supervises the generator to produce high-fidelity speech sample. The min-max objective can be expressed as:

$$\min_\theta \sum_{t\geq 1} \mathbb{E}_{q(t)}\left[D_{\text{adv}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)\right], \tag{4}$$

$$\mathcal{L}_G = \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\left[\left(D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right) - 1\right)^2\right], \tag{5}$$

$$\mathcal{L}_D = \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_t)}\left[\left(D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right) - 1\right)^2\right] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\left[D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)^2\right], \tag{6}$$

Where D_{adv} depends on the adversarial training setup, and the fake samples from $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ are contrasted against the real one from $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$.

Reparameterization on diffusion model. Different from the conventional diffusion models that require hundreds of steps with small β_t to estimate the gradient for data density, recent works (Salimans & Ho, 2022; Liu et al., 2022) have witnessed that approximating some surrogate variables, e.g., the noiseless target data gives better quality. We reparameterize the denoising model by directly predicting the clean data $\mathbf{x}_0$. Free from estimating the gradient for data density, it only needs to predict unperturbed $\mathbf{x}_0$ and then add perturbation with the posterior distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ (formulated in Appendix B), and the reverse transition distribution can be expressed as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t|t, c)\right) \tag{7}$$

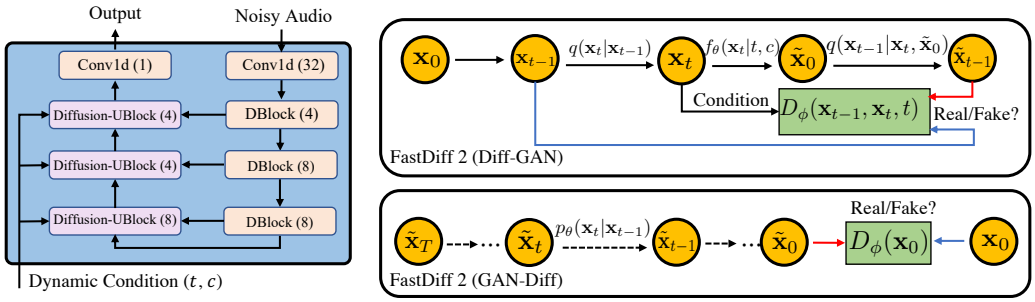
4.3 GAN LEVERAGING DIFFUSION MECHANISM

GAN-based models are often difficult to train, collapsing (Mao et al., 2019) without carefully selected hyperparameters and regularizers, and showing less sample diversity. Besides, these models show distinct degradation in training stability, which cannot generate deterministic values due to the complex data distribution.

In this work, we propose **FastDiff 2 (GAN-Diff)** leveraging diffusion mechanism to construct the forward process by multiple denoising iterations, and thus we expect it exhibits better training stability compared to traditional one-shot GAN. To be more specific, we 1) initialize the generator G with a pre-trained FastDiff teacher; 2) conduct 4-iter denoising to generate $\tilde{\mathbf{x}}_0$ with gradient, which is regarded to be the forward process of the generator; and finally 3) G plays an adversarial game with the discriminator D, and the min-max objective can be expressed as:

$$\mathcal{L}_G = \mathbb{E}_{q_{data}}\left[\left(D_\phi\left(\tilde{\mathbf{x}_0}\right) - 1\right)^2\right] \tag{8}$$

$$\mathcal{L}_D = \mathbb{E}_{q_{data}}\left[\left(D_\phi\left(\tilde{\mathbf{x}_0}\right)\right)^2 + \left(D_\phi\left(\mathbf{x}_0 - 1\right)\right)^2\right] \tag{9}$$



(a) FastDiff 2 architecture (b) Training process of FastDiff 2

Figure 2: The overall architecture for FastDiff 2. In subfigure (a), FastDiff 2 takes noisy audio $\mathbf{x}_t$ as input and conditions on diffusion time index t and mel-spectrogram c.

By distilling the behavior of the converged diffusion teacher (FastDiff) into a new model, it reduces the difficulties of adversarial learning by orders of magnitude. FastDiff 2 (GAN-Diff) breaks the forward process of one-shot conditional GAN into several denoising diffusion iterations, in which each step is relatively simple to model. Thus, it exhibits better training stability and avoids mode collapse.

4.4 Architecture

As illustrated in Figure 2(a), we take FastDiff (Huang et al., 2022) as the model backbone, which includes a stack of time-aware location-variable convolution with diverse receptive field patterns to model long-term time dependencies with adaptive conditions efficiently. Convolution is conditioned on dynamic variations (diffusion steps and spectrogram fluctuations) in speech, which equips the model with diverse receptive field patterns and promotes the robustness of diffusion models.

We build the basic architecture of discriminator upon WaveNet (Oord et al., 2016), which is one of the most popular speech backbones. It consists of ten layers of non-causal dilated 1-D convolutions with weight normalization. The discriminator is trained to correctly classify the generated sample as fake while classifying the ground truth as real. More details have been attached in Appendix C.

4.5 Loss Objective

Adversarial GAN Objective For the generator and discriminator, the training objectives follow (Mao et al., 2017), which replace the binary cross-entropy terms of the original GAN objectives (Goodfellow et al., 2014) with least squares loss functions for non-vanishing gradient flows.

Frequency-domain Reconstruction Objective To stabilize auxiliary learning, we include frequency-domain sample reconstruction loss objective by applying the multi-resolution STFT (Short Time Fourier Transform) operation $STFT(\cdot)$ (given in Appendix F):

$$\mathcal{L}_\theta = \mathcal{L}_{STFT}(\tilde{\boldsymbol{x}}_0, \boldsymbol{x}_0) \tag{10}$$

4.6 Training Algorithm

The training procedures of the proposed FastDiff 2 (GAN-Diff) and FastDiff 2 (Diff-GAN) have been illustrated as follows. The sampling algorithms have been attached in Appendix D.2.

Algorithm 1 Training FastDiff 2 (Diff-GAN)

1: **Require**: FastDiff 2 (Diff-GAN) generator θ, discriminator ϕ, and mel condition c.
2: **repeat**
3: Sample $\mathbf{x}_0 \sim q_{\text{data}}$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $t \sim \text{Unif}(\{1, \cdots, T\})$
4: Sample $\mathbf{x}_t$, $\mathbf{x}_{t-1}$ according to E.q (2)
5: $\tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t|t, c)$
6: Sample $\tilde{\mathbf{x}}_{t-1} \sim q(\mathbf{x}_{t-1}|\mathbf{x}_t, \tilde{\mathbf{x}}_0)$ according to E.q (7)
7: Take gradient descent steps on $\nabla_\theta(\mathcal{L}_\theta + \mathcal{L}_G)$ according to E.q (10) and (5)
8: Take gradient descent steps on $\nabla_\phi \mathcal{L}_D$ according to E.q (6)
9: **until** FastDiff 2 (Diff-GAN) converged

Algorithm 2 Training FastDiff 2 (GAN-Diff)

Require: FastDiff teacher α with schedule β ($T = 4$) derived by noise predictor, FastDiff 2 (GAN-Diff) generator θ, discriminator ϕ, and mel condition c.
Initialize θ parameters using teacher α
repeat
  for $t = T, \cdots, 1$ **do**
    Sample $\tilde{\mathbf{x}}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$
  end for
  Take gradient descent steps on $\nabla_\theta(\mathcal{L}_\theta + \mathcal{L}_G)$ according to E.q (10) and (8)
  Take gradient descent steps on $\nabla_\phi \mathcal{L}_D$ according to E.q (9)
until FastDiff 2 (GAN-Diff) converged

5 Related Works

5.1 Diffusion Probabilistic Model

The diffusion probabilistic model is a family of generative models with the capacity to learn complex data distribution, which has recently attracted a lot of research attention in several important domains, including image synthesis (Dhariwal & Nichol, 2021; Xiao et al., 2021), audio synthesis (Lam et al., 2022; Huang et al., 2022), and 3D point cloud generation (Luo & Hu, 2021). These diffusion-based models can generate high-fidelity samples yet inherently suffer from slow sampling speed. Multiple methods have conducted extensive investigations to accelerate the sampling process: Chen et al. (2020) utilize a grid search algorithm for a shorter inference schedule. Liu et al. (2021) introduces a shallow diffusion mechanism that starts denoising at a particular distribution instead of Gaussian white noise. Xiao et al. (2021) tackle the generative learning trilemma with denoising diffusion GANs, and Liu et al. (2022) propose a denoising diffusion generative adversarial network to achieve high-fidelity and efficient text-to-spectrogram synthesis. However, these works mainly focus on image generation, and we aim to design diffusion models for fast conditional speech synthesis while maintaining leading performance, which has been relatively overlooked.

5.2 Generative Adversarial Network

Generative adversarial networks (GANs) (Jang et al., 2021; Kong et al., 2020a) are one of the most dominant deep generative models for speech generation. UnivNet (Jang et al., 2021) has demonstrated its success in capturing different waveform intervals with local-variable convolution. HIFI-GAN (Kong et al., 2020a) proposes multi-receptive field fusion (MRF) to model the periodic patterns matters. However, GAN-based models are often difficult to train, collapsing (Creswell et al., 2018) without carefully selected hyperparameters and regularizers, and showing less sample diversity. Differently, we incorporate GANs into diffusion models which break the generation process into several conditional denoising steps, in which each step is relatively simple to model. Thus, we expect our model to exhibit better training stability and mode coverage.

6 Experiments

6.1 Experimental Setup

6.1.1 Dataset

For a fair and reproducible comparison against other competing methods, we used the benchmark LJSpeech dataset (Ito & Johnson, 2017) which consists of 13,100 audio clips of 22050 Hz from a female speaker for about 24 hours. To evaluate the model generalization ability over unseen speakers in multi-speaker scenarios, we prepare the VCTK dataset (Yamagishi et al., 2019), which is downsampled to 22050 Hz to match the sampling rate with the LJSpeech dataset. VCTK consists of approximately 44,200 audio clips uttered by 109 native English speakers with various accents. Following the common practice, we conduct preprocessing and extract the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples.

Table 2: Comparison with other neural vocoders in terms of quality and synthesis speed. For sampling, we used 50 steps in WaveGrad, 6 steps in DiffWave and 4 steps in FastDiff, respectively, following ivanvovk (2020), philsyn (2021), and Huang (2022).

Model	MOS (↑)	STOI (↑)	PESQ (↑)	RTF (↓)
GT	4.32±0.06	/	/	/
WaveNet (MOL)	3.95±0.08	/	/	85.230
WaveGlow	3.86±0.08	0.961	3.20	0.029
HIFI-GAN	3.98±0.10	0.954	3.23	0.002
UnivNet	4.05±0.10	0.969	3.54	0.002
Diffwave (6 steps)	4.06±0.09	0.966	3.72	0.093
WaveGrad (50 steps)	4.00±0.00	0.954	3.33	0.390
FastDiff (4 steps)	4.09±0.10	0.971	3.78	0.017
FastDiff 2 (Diff-GAN) (4 steps)	**4.16±0.10**	0.972	3.73	0.017
FastDiff 2 (GAN-Diff) (4 steps)	4.12±0.08	**0.979**	**3.90**	0.017

6.1.2 Model Configurations

FastDiff 2 comprises three Diffusion-UBlocks and DBlocks with the up/downsample rate of $[8, 8, 4]$, respectively. The discriminator consists of ten layers of non-causal dilated 1-D convolutions, whose strides are linearly increasing from one to eight except for the first and last layers. Channels and kernel sizes are set to 64 and 5, respectively. Both variants of FastDiff 2 share the same number of denoising steps ($T = 4$) in both training and inference. The multi-resolution STFT loss is computed by the sum of three different STFT losses described in Appendix F.

6.1.3 Training and Evaluation

Both variants of FastDiff 2 are trained with constant learning rate $lr = 2 \times 10^{-4}$ on 4 NVIDIA V100 GPUs. We use random short audio clips of 25600 samples from each utterance with a batch size of 16 for each GPU. We crowd-source 5-scale MOS tests via Amazon Mechanical Turk to evaluate the audio quality. The MOS scores are recorded with 95% confidence intervals (CI). Raters listen to the test samples randomly, where they are allowed to evaluate each audio sample once. We adopt additional objective evaluation metrics including STOI (Taal et al., 2010) and PESQ (Rix et al., 2001) to test sample quality. To evaluate the sampling speed, we implement the real-time factor (RTF) assessment on a single NVIDIA V100 GPU. More information about objective and subjective evaluation is attached in Appendix E.

6.2 Comparsion With Other Models

We compared our proposed FastDiff 2 in audio quality and sampling speed with competing models, including 1) WaveNet (Oord et al., 2016), the autoregressive generative model for raw audio. 2) WaveGlow (Prenger et al., 2019), the parallel flow-based model. 3) HIFI-GAN V1 (Kong et al., 2020a) and UnivNet (Jang et al., 2021), the most popular GAN-based models. 4) Diffwave (Kong et al., 2020b), WaveGrad (Chen et al., 2020), and FastDiff (Huang et al., 2022), three diffusion probabilistic models that generate high-fidelity speech samples. For easy comparison, the results are compiled and presented in Table 2, and we have the following observations:

In terms of audio quality, FastDiff 2 (Diff-GAN) achieves the highest MOS with scores of 4.16 (Diff-GAN) and 4.12 (GAN-Diff) compared with the baseline models, with a gap of 0.16 compared to the ground truth audio. For objective evaluation, FastDiff 2 also demonstrates the outperformed performance in PESQ and STOI, superior to all baseline models. Regarding inference speed, FastDiff 2 enjoys an effective 4-iter sampling process as FastDiff and enables a speed of 58x faster than real-time on a single NVIDIA V100 GPU without engineered kernels, making diffusion models more practically applicable compared with competing diffusion architectures.

6.3 Ablation Study

We conduct ablation studies to demonstrate the effectiveness of several designs, including the diffusion reparameterization and frequency-domain reconstruction objective in FastDiff 2. The results

of both subjective and objective evaluation have been presented in Table 3, and we have the following observations: 1) Replacing the diffusion reparameterization design and parameterizing the denoising model by predicting the Gaussian noise ϵ has witnessed a distinct degradation in perceptual quality. Specifically, FastDiff 2 (Diff-GAN) directly predicts clean data to avoid significant degradation when reducing reverse iterations. 2) Removing the sample reconstruction loss objective results in blurry predictions with distinct artifact (Kumar et al., 2019) in both variants of FastDiff 2, demonstrating the effectiveness of the auxiliary sample reconstruction regularization in stabilizing adversarial learning.

Table 3: Ablation study results. Comparison of the effect of each component on quality.

Model	MOS (↑)	STOI(↑)	PESQ (↑)
GT	4.32±0.06	/	/
FastDiff 2 (Diff-GAN)	4.16±0.10	**0.972**	3.73
w/o Diffusion Reparameterization	2.40±0.08	0.922	3.19
w/o Reconstruction Objective	2.40±0.08	0.922	3.19
FastDiff 2 (GAN-Diff)	4.12±0.08	**0.979**	**3.90**
w/o Reconstruction Objective	2.71±0.07	0.954	3.15

6.4 Generalization To Unseen Speakers

We use 40 randomly selected utterances of 5 unseen speakers in the VCTK dataset that are not used in training for out-of-distribution testing. Table 4 shows the experimental results for the mel-spectrogram inversion of the samples from unseen speakers: We notice that both variants of FastDiff 2 generate high-fidelity samples and outperform baseline models. In summary, FastDiff 2 universally generates audio with strong robustness from entirely new speakers outside the training set.

Table 4: Comparison with other neural vocoders of synthesized utterances for unseen speakers.

Model	MOS (↑)	STOI(↑)	PESQ (↑)
GT	4.30±0.06	/	/
WaveNet (MOL)	3.80±0.07	/	/
WaveGlow	3.65±0.07	0.870	3.10
HIFI-GAN	3.76±0.09	0.862	3.14
UnivNet	3.79±0.08	0.887	3.21
Diffwave (6 steps)	3.80±0.09	0.873	3.22
WaveGrad (50 steps)	3.73±0.07	0.856	3.15
FastDiff (4 steps)	3.84±0.08	0.894	3.25
FastDiff 2 (Diff-GAN) (4 steps)	**3.96±0.07**	0.910	3.28
FastDiff 2 (GAN-Diff) (4 steps)	3.92±0.08	**0.912**	**3.57**

7 Conclusion

In this work, through investigations of two popular classes (diffusion models and GANs) of deep generative models, we observed that GANs tended to generate samples but did not cover the whole distribution, and the degree of distribution coverage did not distinctly impact audio quality. Inspired by these, we proposed to trade off diversity for quality and speed by incorporating GANs into diffusion models, introducing two GAN-empowered modeling perspectives: 1) FastDiff 2 (Diff-GAN), a diffusion model whose denoising process was parametrized by conditional GANs, and the non-Gaussian denoising distribution made it much more stable to implement the reverse process with large step sizes; and 2) FastDiff 2 (GAN-Diff): a generative adversarial network whose forward process was constructed by multiple denoising diffusion iterations, and it exhibited better training stability and mode coverage. Unlike the acceleration strategies based on jumping denoising steps that only accelerated the generation process during inference, FastDiff 2 provided a principled way to accelerate DDPMs in both training and inference. Experimental results showed that both variants of FastDiff 2 enjoyed an efficient 4-step sampling process as in FastDiff yet demonstrated a superior sample quality. We envisage that our work could serve as a basis for future speech synthesis studies.

REFERENCES

Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. Avocodo: Generative adversarial network for artifact-free vocoder. *arXiv preprint arXiv:2206.13404*, 2022.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *Proc. of ICLR*, 2020.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, volume 34, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.

Huang. Fastdiff. 2022.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017. Accessed: 2022-01-01.

ivanvovk. Wavegrad. 2020.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Proc. of InterSpeech*, 2021.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of NeurIPS*, 33:17022–17033, 2020a.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proc. of ICLR*, 2020b.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

Max WY Lam, Jun Wang, Dan Su, and Dong Yu. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *Proc. of ICLR*, 2022.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *arXiv preprint arXiv:2105.02446*, 2, 2021.

Songxiang Liu, Dan Su, and Dong Yu. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*, 2022.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proc. of CVPR*, 2019.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive gan for conditional waveform synthesis. *arXiv preprint arXiv:2110.10139*, 2021.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

philsyn. Diffwave-vocoder. 2021.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proc. of ICML*, 2021.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. of ICASSP*, 2019.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: fast, robust and controllable text to speech. In *Proc. of ICONIP*, 2019.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. of ICASSP*, 2001.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. of ICLR*, 2020.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. of ICASSP*, 2010.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 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)*, 2019.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. of ICASSP*, 2020.

A DETAILED FORMULATION OF DDPM

We define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov chain from data $\mathbf{x}_0$ to the latent variable $\mathbf{x}_T$:

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{11}$$

For a small positive constant β_t, a small Gaussian noise is added from $\mathbf{x}_t$ to the distribution of $\mathbf{x}_{t-1}$ under the function of $q(\mathbf{x}_t|\mathbf{x}_{t-1})$.

The whole process gradually converts data $\mathbf{x}_0$ to whitened latents $\mathbf{x}_T$ according to the fixed noise schedule $\beta_1, \cdots, \beta_T$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{12}$$

Efficient training is optimizing a random term of t with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2 \tag{13}$$

Unlike the diffusion process, the reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from x_T to x_0 parameterized by shared θ:

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{14}$$

where each iteration eliminates the Gaussian noise added in the diffusion process:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{15}$$

B DIFFUSION POSTERIOR DISTRIBUTION

Firstly we compute the corresponding constants respective to diffusion and reverse process:

$$\alpha_t = \prod_{i=1}^{t} \sqrt{1-\beta_i} \quad \sigma_t = \sqrt{1-\alpha_t^2} \tag{16}$$

The Gaussian posterior in the diffusion process is defined through the Markov chain, where each iteration adds Gaussian noise. Consider the forward diffusion process in Eq. 12, which we repeat here:

$$\begin{aligned} q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|x_0) &= \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \\ q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &= \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}) \end{aligned} \tag{17}$$

We emphasize the property observed by (Ho et al., 2020), the diffusion process can be computed in a closed form:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t \mathbf{I}) \tag{18}$$

C MODEL HYPERPARAMETERS

C.1 ARCHITECTURES

As illustrated in Table 5, we list the hyper-parameters of FastDiff 2.

Table 5: Architecture hyperparameters of FastDiff 2.

Hyperparameter	FastDiff 2
DBlock Hidden Channels	32
DBlock Downsample Ratios	[4, 8, 8]
Diffusion UBlock Hidden Channels	32
Diffusion UBlock Upsample Ratios	[8, 8, 4]
Time-aware LVC layers Each Block	4
Time-aware LVC layers Kernel Size	256
Diffusion Kernel Predictor Hidden Channels	64
Diffusion Kernel Predictor Kernel Size	3
Diffusion Embedding Input Channels	128
Diffusion Embedding Output Channels	512
Use Weight Norm	True
Total Number of Parameters	15 M

Table 6: Diffusion hyperparameters of FastDiff 2.

Diffusion Hyperparameter
FastDiff 2 (GAN-Diff): $\beta = [3.6701e^{-7}, 1.7032e^{-5}, 7.908e^{-4}, 7.6146e^{-1}]$
FastDiff 2 (Diff-GAN): $\beta = \text{Linear}(1 \times 10^{-4}, 0.1, 4)$

C.2 Diffusion hyperparameters

We list the diffusion hyper-parameters of FastDiff 2 in Table 6.

D Training and Inference details

D.1 Preliminary Study

Both models are trained with constant learning rate $lr = 2 \times 10^{-4}$ on 4 NVIDIA V100 GPUs. We conduct preprocessing and extract the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples.

For audio quality, we adopt objective evaluation metrics including MCD (Kubichek, 1993) and PESQ (Rix et al., 2001). We crowd-sourced 5-scale MOS tests via Amazon Mechanical Turk. Raters listened to the test samples randomly, where they were allowed to evaluate each audio sample once. To evaluate the sampling speed, we implement the real-time factor (RTF) assessment on a single NVIDIA V100 GPU. NDB and JSD metrics are employed to explore the diversity of generated mel-spectrograms.

D.2 Sampling Algorithm

Algorithm 3 Sampling with FastDiff 2 (Diff-GAN)

1: **Input**: FastDiff 2 (Diff-GAN) generator θ, and mel condition c.
2: Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
3: **for** $t = T, \cdots, 1$ **do**
4: Sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = q(\mathbf{x}_{t-1}|\mathbf{x}_t, \tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t|t, c))$
5: **end for**
6: **return** $\mathbf{x}_0$

Algorithm 4 Sampling with FastDiff 2 (GAN-Diff)

1: **Input**: FastDiff 2 (GAN-Diff) generator θ, and mel condition c.
2: Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
3: **for** $t = T, \cdots, 1$ **do**
4: Sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
5: **end for**
6: **return** $\mathbf{x}_0$

E Evaluation Matrix

E.1 Objective Evaluation

Perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and The short-time objective intelligibility (STOI) (Taal et al., 2010) assesses the denoising quality for speech enhancement.

Number of Statistically-Different Bins (NDB) and Jensen-Shannon divergence (JSD). They measure diversity by 1) clustering the training data into several clusters, and 2) measuring how well the generated samples fit into those clusters.

Mel-cepstral distortion (MCD) (Kubichek, 1993) measures the spectral distance between the synthesized and reference mel-spectrum features.

E.2 Subjective Evaluation

All our Mean Opinion Score (MOS) tests are crowd-sourced and conducted by native speakers. The scoring criteria have been included in Table 7 for completeness. The samples are presented and rated one at a time by the testers, each tester is asked to evaluate the subjective naturalness of a sentence on a 1-5 Likert scale. The screenshots of instructions for testers are shown in Figure 3. We paid $8 to participants hourly and totally spent about $600 on participant compensation.

Table 7: Ratings that have been used in the evaluation of speech naturalness of synthetic and ground truth samples.

Rating	Naturalness	Definition
1	Bad	Very annoying and objectionable dist.
2	Poor	Annoying but not objectionable dist.
3	Fair	Perceptible and slightly annoying dist
4	Good	Just perceptible but not annoying dist.
5	Excellent	Imperceptible distortions

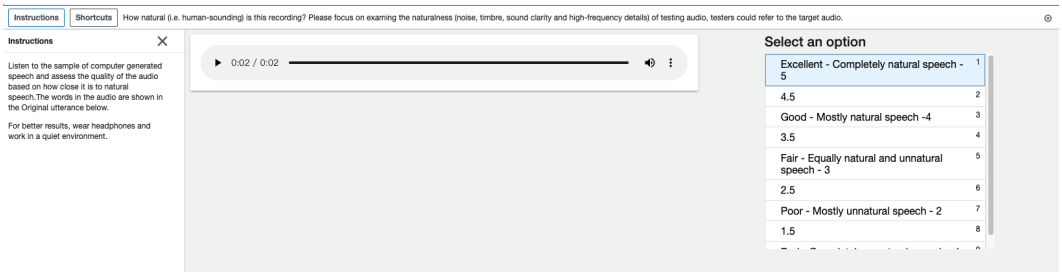


Figure 3: Screenshot of MOS testing.

F MULTI-RESOLUTION STFT LOSS DETAILS

By applying the multi-resolution short time fourier transform, we respectively obtain the spectral convergence ($\mathcal{L}_{stft\_sc}$) and log STFT magnitude ($\mathcal{L}_{stft\_mag}$) of $\mathcal{L}_{STFT}$ in frequency domain:

$$\mathcal{L}_{stft\_sc} = \frac{\| \operatorname{STFT}(\mathbf{x}_0) - \operatorname{STFT}(\tilde{\mathbf{x}}_0) \|_F}{\| \operatorname{STFT}(\mathbf{x}_0) \|_F} \tag{19}$$

$$\mathcal{L}_{stft\_mag} = \frac{1}{N} \| \log(\operatorname{STFT}(\mathbf{x}_0)) - \log(\operatorname{STFT}(\tilde{\mathbf{x}}_0)) \|_1, \tag{20}$$

where $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius and L1 norms. N denotes the number of elements in the magnitude; The final multi-resolution STFT loss is the sum of M losses with different analysis parameters(i.e., FFT size, window size, and hop size), and we set $M = 3$:

$$\mathcal{L}_{STFT} = \frac{1}{M} \sum_{m=1}^{M} \left(\mathcal{L}_{stft\_sc}^{(m)} + \mathcal{L}_{stft\_mag}^{(m)} \right) \tag{21}$$

Table 8: The details of the multi-resolution STFT loss. A Hanning window was applied before the FFT process.

FFT size	Frame shift	Window size
1024	600	120
2048	120	250
512	240	50