

FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis

Rongjie Huang^{1*†}, Max W. Y. Lam^{2†}, Jun Wang², Dan Su², Dong Yu³, Yi Ren¹, Zhou Zhao^{1‡}

¹Zhejiang University

²Tencent AI Lab, China

³Tencent AI Lab, USA

{rongjiehuang, rayeren, zhouzhao}@zju.edu.cn, {maxwylam, joinerwang, dansu, dyu}@tencent.com

Abstract

Denoising diffusion probabilistic models (DDPMs) have recently achieved leading performances in many generative tasks. However, the inherited iterative sampling process costs hindered their applications to speech synthesis. This paper proposes FastDiff, a fast conditional diffusion model for high-quality speech synthesis. FastDiff employs a stack of time-aware location-variable convolutions of diverse receptive field patterns to efficiently model long-term time dependencies with adaptive conditions. A noise schedule predictor is also adopted to reduce the sampling steps without sacrificing the generation quality. Based on FastDiff, we design an end-to-end text-to-speech synthesizer, FastDiff-TTS, which generates high-fidelity speech waveforms without any intermediate feature (e.g., Mel-spectrogram). Our evaluation of FastDiff demonstrates the state-of-the-art results with higher-quality (MOS 4.28) speech samples. Also, FastDiff enables a sampling speed of 58x faster than real-time on a V100 GPU, making diffusion models practically applicable to speech synthesis deployment for the first time. We further show that FastDiff generalized well to the mel-spectrogram inversion of unseen speakers, and FastDiff-TTS outperformed other competing methods in end-to-end text-to-speech synthesis. Audio samples are available at <https://FastDiff.github.io/>.

1 Introduction

With the recent development of deep generative models, speech synthesis has seen an extraordinary progress. Among the conventional speech synthesis methods, WaveNets [Oord *et al.*, 2016] were demonstrated to generate high-fidelity audio samples in an autoregressive manner yet suffering from prohibitively expensive computational costs. In contrast, non-autoregressive approaches such as flow-based and GAN-based models [Prenger *et al.*, 2019; Jang *et al.*, 2021;

Kong *et al.*, 2020a; Huang *et al.*, 2021] were also proposed to generate speech audios with satisfactory speed. However, these models were still criticized for other problems, e.g., the limited sample quality or sample diversity [Xiao *et al.*, 2021].

In speech synthesis, our goal is mainly two-fold:

- High-quality: generating high-quality speech is a challenging problem especially when the sampling rate of an audio is high. It is vital to reconstruct details at different timescales for waveforms of highly variable patterns.
- Fast: high generation speed is essential when considering real-time speech synthesis. This poses a challenge for all high-quality neural synthesizers.

As a blossoming class of generative models, denoising diffusion probabilistic models (DDPMs) [Ho *et al.*, 2020; Song *et al.*, 2020; Lam *et al.*, 2022] has emerged to prove its capability to achieve leading performances in both image and audio syntheses [Dhariwal and Nichol, 2021; Kong *et al.*, 2020b; Chen *et al.*, 2020; Lam *et al.*, 2022]. However, current development of DDPMs in speech synthesis was hampered by two major challenges:

- Different from other existing generative models, diffusion models are not trained to directly minimize the difference between the generated audio and the reference audio, but to de-noise a noisy sample given an optimal gradient. This in practice could lead to overly de-noised speech after a large number of sampling steps, in which natural voice characteristics including breathiness and vocal fold closure are removed.
- While DDPMs inherently are gradient-based models, a guarantee of high sample quality typically comes at a cost of hundreds to thousands of de-noising steps. When reducing the sampling steps, an apparent degradation in quality due to perceivable background noise is observed.

In this work, we propose FastDiff, a fast conditional diffusion model for high-quality speech synthesis. To improve audio quality, FastDiff adopts a stack of time-aware location-variable convolutions of diverse receptive field patterns to efficiently model long-term time dependencies with adaptive conditions. To accelerate the inference procedure, FastDiff also includes a noise schedule predictor, which derives a short and effective noise schedule and significantly reduces the denoising steps. Based on FastDiff, we also introduce an end-to-end phoneme-to-waveform synthesizer FastDiff-TTS, which

*Work done during internship at Tencent AI Lab

†Equal contribution

‡Corresponding author

simplifies the text-to-speech generation pipeline and does not require intermediate features or specialized loss functions to enjoy low inference latency.

Experimental results demonstrated that FastDiff achieved a higher MOS score than the best publicly available models and outperformed the strong WaveNet vocoder (MOS: 4.28 vs. 4.20). FastDiff further enjoys an effective sampling process and only needs 4 iterations to synthesize high-fidelity speech, 58x faster than real-time on a V100 GPU without engineered kernels. To the best of our knowledge, FastDiff is the first diffusion model with a sampling speed comparable to previous for the first time applicable to interactive, real-world speech synthesis applications at a low computational cost. FastDiff-TTS successfully simplify the text-to-speech generation pipeline and outperform competing architectures.

2 Background: Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) [Ho *et al.*, 2020; Song *et al.*, 2020; Lam *et al.*, 2022] are likelihood-based generative models that have recently succeeded to advance the state-of-the-art results in benchmark generative tasks [Dhariwal and Nichol, 2021] and have proved its capability to produce high-quality samples. The basic idea of DDPMs is to train a gradient neural network for reversing a diffusion process. Given i.i.d. samples $\{x_0 \in \mathbb{R}^D\}$ from an unknown data distribution $p_{data}(x_0)$, DDPMs try to approximate $p_{data}(x_0)$ by a marginal distribution $p_\theta(x_0) = \int \cdots \int p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) dx_1 \dots dx_T$.

In data distribution as $q(x_0)$, the diffusion process is defined by a fixed Markov chain from data x_0 to the latent variable x_T . For a small positive constant β_t , a small Gaussian noise is added from x_t to the distribution of x_{t-1} under the function of $q(x_t|x_{t-1})$. The whole process gradually converts data x_0 to whitened latents x_T according to the fixed noise schedule β_1, \dots, β_T . The reverse process is a Markov chain from x_T to x_0 parameterized by a shared θ , which aims to recover samples from Gaussian noises through eliminating the Gaussian noise added in the diffusion process in each iteration.

It has been demonstrated that diffusion probabilistic models [Dhariwal and Nichol, 2021; Xiao *et al.*, 2021] can learn diverse data distribution in multiple domains, such as images and time series. While the main issue with the proposed neural diffusion process is that it requires up to thousands of iterative steps to reconstruct the target distribution during reverse sampling. In this work, we offer a fast conditional diffusion model to reduce reverse iterations and improve computational efficiency.

3 FastDiff

This section presents our proposed FastDiff, a fast conditional diffusion model for high-quality speech synthesis. We first describe the motivation of the design in FastDiff. Secondly, we introduce the iterative refinement model θ for high-quality speech synthesis and the noise predictor ϕ for accelerated sampling. Furthermore, we describe the training and

inference procedures in detail. At last, we extend FastDiff to FastDiff-TTS for fully end-to-end text-to-speech syntheses.

3.1 Motivation

While denoising diffusion probabilistic models have shown high potential in synthesizing high-quality speech samples [Chen *et al.*, 2020; Kong *et al.*, 2020b; Liu *et al.*, 2021], several challenges remain for industrial deployment: 1) Different from the traditional generative models, diffusion models catch dynamic dependencies from noisy audio instead of clean ones, which introduce more variation information (i.e., noise levels) in addition to the spectrogram fluctuation. 2) With limited receptive field patterns, a distinct degradation could emerge when reducing the reverse iterations, making diffusion models difficult to get accelerated. As a result, hundred or thousand orders of iterations prevent existing diffusion models from real-world deployment.

In FastDiff, we propose two key components to complement the above issues: 1) FastDiff adopts a time-aware location-variable convolution to catch the details of noisy samples at dynamic dependencies. The convolution operations are conditioned on dynamic variations in speech including diffusion steps and spectrogram fluctuations, equipping the model with diverse receptive field patterns and promoting the robustness of diffusion models during reverse acceleration. 2) To accelerate the inference procedure, FastDiff adopts a noise schedule predictor to reduce the number of reverse iterations, frees diffusion models from hundreds or thousands of refinement iterations. This makes FastDiff for the first time applicable to interactive, real-world applications at a low computational cost.

3.2 Time-Aware Location-Variable Convolution

In comparison with traditional convolution networks, location-variable convolution [Zeng *et al.*, 2021] shows efficiency in modeling the long-term dependency of audio and gets neural network free from a significant number of dilated convolution layers. Inspired by this, we introduce the Time-Aware Location-Variable Convolution, which is sensitive to time steps in diffusion probabilistic models. At time step t , we follow [Vaswani *et al.*, 2017] to embed the step index into an 128-dimensional positional encoding (PE) vector e_t :

$$e_t = \left[\sin \left(10^{\frac{0 \times 4}{63}} t \right), \dots, \sin \left(10^{\frac{63 \times 4}{63}} t \right), \right. \\ \left. \cos \left(10^{\frac{0 \times 4}{63}} t \right), \dots, \cos \left(10^{\frac{63 \times 4}{63}} t \right) \right],$$

In time-aware location-variable convolution, FastDiff requires multiple predicted variation-sensitive kernels to perform convolutional operations on the associated intervals of input sequence. These kernels should be time-aware and sensitive to variations of noisy audio including diffusion steps and acoustic features (i.e., Mel-spectrogram). Therefore, we propose a time-aware location-variable convolution (LVC) module, which is coupled with a kernel predictor as shown in Figure 1(b) and Figure 1(c). We describe the overall calculations below.

For the q -th time-aware LVC layer, we split the input $x_t \in \mathbb{R}^D$ using a M -length window with 3^q dilations to produce K

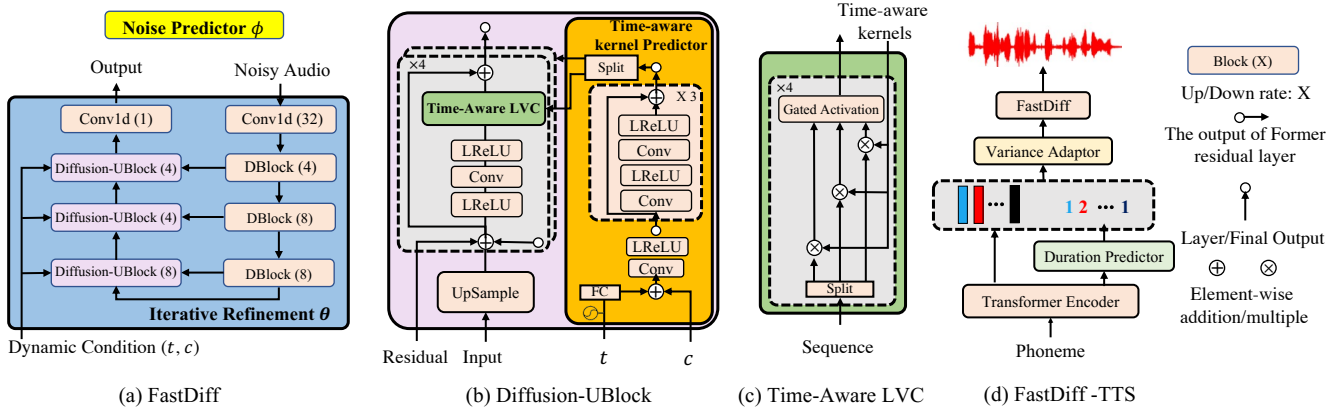


Figure 1: The overall architecture for FastDiff and FastDiff-TTS. The refinement model θ in FastDiff takes noisy audio \mathbf{x}_t as input and computes $\epsilon_\theta(\mathbf{x}_t|c, t)$ conditioned on diffusion time index t and Mel-spectrogram c . We use LReLU to denote the leaky rectified linear unit, LVC to denote the location-variable convolution, FC to denote the fully-connected layer, and PE to denote the positional encoding operation.

segments with each $\mathbf{x}_t^k \in \mathbb{R}^M$:

$$\{\mathbf{x}_t^1, \dots, \mathbf{x}_t^K\} = \text{split}(\mathbf{x}_t; M, q) \quad (1)$$

Next, we perform convolutional operations on the associated intervals of input sequence using the kernels generated by a kernel predictor α :

$$\{\mathbf{F}_t, \mathbf{G}_t\} = \alpha(t, c) \quad (2)$$

$$\mathbf{z}_t^k = \tanh(\mathbf{F}_t * \mathbf{x}_t^k) \odot \sigma(\mathbf{G}_t * \mathbf{x}_t^k) \quad (3)$$

$$\mathbf{z}_t = \text{concat}(\{\mathbf{z}_t^1, \dots, \mathbf{z}_t^K\}), \quad (4)$$

where $\mathbf{F}_t, \mathbf{G}_t$ denote the filter and the gate kernels for \mathbf{x}_t^i , respectively, $*$ denotes the 1d convolution, \odot denotes the element-wise product and $\text{concat}(\cdot)$ denotes the concatenation between vectors. Since the time-aware kernels are adaptive to the noise-level and dependent to the acoustic features, FastDiff is capable of precisely estimating de-noising gradient with a superior speed given a noisy signal input.

3.3 Accelerated Sampling

Noise Predictor

To avoid sampling with hundreds to thousands steps, FastDiff adopts the noise scheduling algorithm in the bilateral denoising diffusion models (BDDMs) [Lam *et al.*, 2022] to predict a sampling noise schedule much shorter than the noise schedule used in training. This scheduling method has been revealed to be superior than other sampling acceleration methods, e.g., the grid search algorithm in WaveGrad [Chen *et al.*, 2020] and the fast sampling algorithm in DiffWave [Kong *et al.*, 2020b]. The noise predictor iteratively derives a continuous noise schedule $\hat{\beta} \in \mathbb{R}^{T_m}$. We attach the learning objective and corresponding likelihood in Appendix.

Schedule Alignment

In FastDiff, similar to DDPMs, during training we use $T = 1000$ discrete time steps. Therefore, when needed to condition on t during sampling, we also need to approximate T_m discrete time indices by aligning the T_m -step sampling noise schedule $\hat{\beta}$ to the T -step training noise schedule β , with $N \ll T$. We have attached the detailed algorithms in Appendix.

3.4 Training, Noise Scheduling and Sampling

All illustrated in Algorithm 1, we separately parameterize FastDiff by two modules: 1) a iterative refinement model θ that minimizes a variational bound of the score function, and 2) a noise predictor ϕ that optimizes the noise schedule for a tighter evidence lower bound. For inference, we first derive the tighter and more efficient noise schedules $\hat{\beta}$ via an one-shot noise scheduling procedure, which makes FastDiff achieve orders of magnitude faster at sampling. It has been demonstrated [Lam *et al.*, 2022] that the noise schedule searched for as few as 1 sample could be robust enough to maintain a high-quality generation among all samples in testing set. Secondly, we map the continuous noise schedules to discrete time indexes T_m using schedule alignment. Finally, FastDiff iteratively refines gaussian noise to generate high-quality samples with computational efficiency. The detailed information on training, noise scheduling and inference procedures has been presented in Appendix.

3.5 FastDiff-TTS

Existing text-to-speech methods usually adopt a two-stage pipeline: 1) A text-to-spectrogram generation module (a.k.a. acoustic model) aims to generate prosodic attributes according to variance prediction; 2) A conditional waveform generation module (a.k.a. vocoder) adds the phase information and synthesizes a detailed waveform. To further simplify the text-to-speech synthesis pipeline, we propose a fully end-to-end model FastDiff-TTS, which does not require intermediate features or specialized loss functions. FastDiff-TTS is designed to be a fully differentiable and efficient architecture that directly produces waveforms from contexts (e.g. phonemes) without needing to generate acoustic features (e.g., Mel-spectrograms) explicitly.

Architecture

The architecture design of FastDiff-TTS refers to a convolutional non-autoregressive TTS model – FastSpeech 2 [Ren *et al.*, 2020] as the backbone. The architecture of FastDiff-TTS is illustrated in Figure 1(d). In FastDiff-TTS, the encoder first converts the phoneme embedding sequence into the phoneme

Algorithm 1 Training refinement network θ

- 1: **Input:** Pre-defined noise schedule β
 - 2: **repeat**
 - 3: Sample $\mathbf{x}_0 \sim q_{data}$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $t \sim \text{Unif}(\{1, \dots, T\})$
 - 4: $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sqrt{1 - \alpha_t^2} \epsilon$
 - 5: Take gradient descent steps on $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\mathbf{x}_t|c, t)\|_2^2$
 - 6: **until** refinement model θ converged
-

Algorithm 2 Training noise predictor ϕ

- 1: **Input:** Pre-defined discrete β , trained refinement network θ , hyperparameter τ .
 - 2: **repeat**
 - 3: Sample $\mathbf{x}_0 \sim q_{data}$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $t \sim \text{Unif}(\{\tau, \dots, T - \tau\})$
 - 4: $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sqrt{1 - \alpha_t^2} \epsilon$
 - 5: $\hat{\beta}_t = \min \left\{ 1 - \alpha_t^2, 1 - \frac{\alpha_{t+\tau}^2}{\alpha_t^2} \right\} \phi(\mathbf{x}_t)$
 - 6: Take gradient descent steps on $\nabla_{\phi} \left\{ \frac{\delta_t^2}{2(\delta_t^2 - \hat{\beta}_t)} \|\epsilon - \frac{\hat{\beta}_t}{\delta_t^2} \epsilon_{\theta}(\mathbf{x}_t|c, t)\|_2^2 \right\}$
 - 7: **until** noise predictor ϕ converged
-

hidden sequence. Then, the duration predictor expands the encoder output to match the length of the desired waveform output. Given the aligned sequence, the variance adaptor adds pitch information into the hidden sequence. Note that it is difficult to use the full audio corresponding to the full text sequence for training due to the typically high sampling rate for high-fidelity waveform and the limited GPU memory. Therefore, we sample a small segment to synthesize the waveform before passing to the FastDiff model. Finally, the FastDiff model decodes the adapted hidden sequence into a speech waveform as in the vocoder task.

Training Loss

FastDiff-TTS does not require specialized loss functions and adversarial training to improve sample quality as suggested by the previous works [Ren *et al.*, 2020; Donahue *et al.*, 2020; Kim *et al.*, 2021]. This, to a large extent, simplifies the text-to-speech generation. The final training loss consists of the following terms: 1) a duration prediction loss L_{dur} : the mean squared error between the predicted and the ground-truth word-level duration in log-scale, 2) a diffusion loss L_{diff} : the mean squared error between the estimated and gaussian noise, and 3) a pitch reconstruction loss L_{pitch} : the mean squared error between the predicted and the ground-truth pitch sequences. We empirically found that the pitch reconstruction loss L_{pitch} is helpful for handling the one-to-many mapping issue in text-to-speech synthesis.

4 Related Works

Text-to-speech (TTS) systems aim to synthesize raw speech waveforms from given text. In recent years, Neural network based TTS [Ren *et al.*, 2020; Kim *et al.*, 2020; Liu *et al.*,

Algorithm 3 Sampling

- 1: **Input:** Searched $\hat{\beta}$ in noise scheduling process.
 - 2: Compute discrete steps T_m sequences via schedule alignment in Section 3.3.
 - 3: Sample $\mathbf{x}_{T_m} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 4: **for** $t = T_m, \dots, 1$ **do**
 - 5: Sample $\mathbf{x}_{t-1} \sim p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t; \hat{\beta})$
 - 6: **end for**
 - 7: **return** \mathbf{x}_0
-

2021] has made huge progress and attracted a lot of attention in the machine learning and speech community.

Neural vocoder plays the most important role in the recent success of speech synthesis, which require diverse receptive field patterns to catch audio dependencies: 1) autoregressive model WaveNet [Oord *et al.*, 2016] requires causal convolutions layers to increase the receptive field while suffering from slow inference speed. 2) Flow-based generative models [Prenger *et al.*, 2019] fully utilize modern parallel computing processors to broaden receptive fields and speed-up sampling, while they usually achieve a limited sample quality. 3) Generative adversarial networks (GANs) [Jang *et al.*, 2021; Kong *et al.*, 2020a] are one of the most dominant deep generative models in audio generation. UnivNet [Jang *et al.*, 2021] has demonstrated its success in using local-variable convolution on different waveform intervals, and 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. 4) Recently proposed diffusion models Diffwave [Kong *et al.*, 2020b] and WaveGrad [Chen *et al.*, 2020] could generate high-quality speech samples, while suffering from a distinct degradation when reducing reverse iterations, making diffusion models difficult to get accelerated.

Another important line of work covers directly waveform generation from text: FastSpeech 2s [Ren *et al.*, 2020] and VITS [Kim *et al.*, 2021] adopt adversarial training process and spectral losses for improving audio quality, while they do not take full advantage of end-to-end training. Recently proposed WaveGrad 2 [Chen *et al.*, 2021] estimates the gradient of the log conditional density of the waveform given a phoneme sequence, but suffers from a large model footprint and slow inference. Unlike all of the aforementioned methods, as highlighted in section 3.5, FastDiff-TTS is a fully differentiable and efficient architecture that produces waveforms directly without generating middle features (e.g., spectrograms) explicitly. In additional, our diffusion probabilistic model gets free from hundred or thousands of iterations and enjoy computational efficiency.

5 Experiments

5.1 Setup

Dataset

For a fair and reproducible comparison against other competing methods, we used the benchmark LJSpeech dataset [Ito

and Johnson, 2017]. LJSpeech consists of 13,100 audio clips of 22050 Hz from a Female speaker with about 24 hours in total. To evaluate the generalization ability of our model over unseen speakers in multi-speaker scenarios, we also used the VCTK dataset [Yamagishi *et al.*, 2019], which was 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. For both datasets, we used 80-band Mel-spectrograms as the condition for the vocoding task. The FFT size, window size, and hop size were, respectively, set to 1024, 1024, and 256.

Model Configurations

FastDiff mainly consists of the refinement model θ and noise schedule predictor ϕ . The refinement model θ comprises three Diffusion-UBlock and DBlock with the upsample or downsample rate of [8, 8, 4], respectively. We adopt a lightweight GALR network effective in separating the added gaussian noise from audio as the noise schedule predictor ϕ . For end-to-end text-to-speech generation, FastDiff-TTS follows the basic structure in FastSpeech 2 [Ren *et al.*, 2020], which consists of 4 feed-forward transformer blocks in the phoneme encoder. More details have been attached in the Appendix.

Training and Evaluation

The complete training pipeline has been illustrated in Algorithm 1: FastDiff was trained with constant learning rate $lr = 2 \times 10^{-4}$. The refinement model θ and noise predictor ϕ were trained for 1M and 10K steps until convergence, respectively. FastDiff-TTS was trained until 500k steps using the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$. Both models were trained on 4 NVIDIA V100 GPUs using random short audio clips of 16,000 samples from each utterance with a batch size of 16 each GPU. More details have been attached in the Appendix.

We crowd-sourced 5-scale MOS tests via Amazon Mechanical Turk to evaluate the audio quality. The MOS scores were recorded with 95% confidence intervals (CI). Raters listened to the test samples randomly, where they were allowed to evaluate each audio sample once. We further include 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 implemented real-time factor (RTF) assessment on a single NVIDIA V100 GPU. In addition, we employed two metrics NDB and JSD to explore the diversity of generated mel-spectrograms. More information about both objective and subjective evaluation has been attached in Appendix.

5.2 Comparison With Other Models

We compared our FastDiff in audio quality, diversity, 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], non-autoregressive flow-based model. 3) HIFI-GAN V1 [Kong *et al.*, 2020a] and UnivNet [Jang *et al.*, 2021], the most dominant and popular GAN-based models. 4) Diffwave [Kong *et al.*, 2020b] and WaveGrad [Chen *et*

al., 2020], recently proposed diffusion probabilistic models which achieve state-of-the-art in speech synthesis. For easy comparison, the results are compiled and presented in Table 1, and we have the following observations:

In terms of audio quality, FastDiff achieved the highest MOS with a gap of 0.24 compared to the ground truth audio, and it matched the performance of the autoregressive WaveNet baseline and outperformed the non-autoregressive baselines. For objective evaluation, FastDiff also demonstrated a large improvement in PESQ and STOI. For inference speed, with the efficient noise schedules searched by noise predictor, FastDiff could generate high-quality speech samples within as few as 4 reverse steps, significantly reducing the inference time compared with competing diffusion architectures. To the best of our knowledge, FastDiff makes diffusion models for the first time applicable to interactive, high-quality real-world speech synthesis at a low computational cost. In terms of sample diversity, we can see that FastDiff still witnessed a gap from autoregressive WaveNet, but it achieve a higher variety for generated speeches than non-autoregressive baselines. More detailed evaluation on sample diversity has been attached in Appendix.

5.3 Ablation Study

We conducted ablation studies to demonstrate the effectiveness of several designs in FastDiff, including the time-aware location variable convolution and noise predictor in neural vocoding. The results of both subjective and objective evaluations have been presented in Table 2, and we have the following observations: 1) Replacing time-aware location-variable convolution by traditional convolutional operations causes a distinct degradation in sampling speed and perceptual quality. 2) Using grid search instead of the noise predictor to search schedules had witnessed the decreased audio quality, demonstrating that the noise schedule prediction process provides more efficient reverse sampling without sacrificing quality.

Further, we compare two variants of FastDiff to test the modality differences of diffusion condition (i.e., continuous noise-level or discrete time-step). Note that the former model does not require the schedule alignment process mentioned in Section 3.3. We empirically find that the FastDiff model conditioned on discrete time steps could synthesize samples with higher quality, demonstrating that learning proposed FastDiff with discrete diffusion times could be a better choice. More information on the variant of FastDiff extended to continuous noise schedules has been attached in Appendix.

5.4 Generalization To Unseen Speakers

We used 50 randomly selected utterances of 5 unseen speakers in the VCTK dataset that were excluded from the training set for the MOS test. Table 3 shows the experimental results for the mel-spectrogram inversion of the unseen speakers: In summary, we noticed that FastDiff achieved state-of-the-art in terms of audio quality for out-of-domain generalization, indicating that FastDiff could universally generate high-fidelity audio from entirely new speakers outside the train set.

Model	Quality			Speed RTF (↓)	Diversity	
	MOS (↑)	STOI (↑)	PESQ (↑)		NDB (↓)	JSD (↓)
GT	4.52±0.09	/	/	/	/	/
WaveNet (MOL)	4.20±0.06	/	/	85.230	33	0.002
WaveGlow	3.89±0.07	0.961	3.16	0.029	66	0.014
HIFI-GAN	4.08±0.08	0.956	3.28	0.002	72	0.010
UnivNet	4.13±0.09	0.971	3.45	0.002	68	0.013
Diffwave (6 steps)	4.18±0.08	0.966	3.62	0.093	72	0.007
WaveGrad (50 steps)	4.09±0.07	0.911	2.70	0.390	61	0.008
FastDiff (4 steps)	4.28±0.07	0.976	3.71	0.017	49	0.006

Table 1: Comparison with other neural vocoders in terms of quality, synthesis speed and sample diversity. For sampling, we used 50 steps in WaveGrad and 6 steps in DiffWave, respectively.

Model	MOS (↑)	STOI(↑)	PESQ (↑)	RTF (↓)
GT	4.52±0.09	/	/	/
w/o Time-aware LVC	4.08±0.05	0.971	3.45	0.081
w/o Noise Predictor	4.10±0.06	0.968	3.50	0.033
Continuous, 4 steps	4.09±0.08	0.970	3.37	0.015
Continuous, 1000 steps	4.14±0.07	0.980	3.64	3.80
Discrete, 4 steps	4.28±0.07	0.976	3.71	0.017
Discrete, 1000 steps	4.36±0.08	0.989	3.86	4.70

Table 2: Ablation study results. Comparison of the effect of each component in terms of quality and synthesis speed.

Model	MOS
GT	4.37±0.06
WaveNet (MOL)	4.01±0.08
WaveGlow	3.66±0.08
HIFI-GAN	3.74±0.06
UnivNet	3.85±0.07
Diffwave (6 steps)	3.90±0.07
WaveGrad (50 steps)	3.72±0.06
FastDiff (4 steps)	4.10±0.06

Table 3: Comparison with other neural vocoders of synthesized utterances for unseen speakers.

5.5 End-to-End Text-to-Speech

To demonstrate the robustness of the proposed model in end-to-end text-to-speech synthesis, we compare FastDiff-TTS with other neural TTS systems, including 1) GT, the ground truth audio; 2) GT (voc.), where we first convert the ground truth audio into mel-spectrograms, and then convert the mel-spectrograms back to audio using FastDiff; 3) PortaSpeech [Ren *et al.*, 2021] + FastDiff: vocoder cascaded with mel-spectrogram generation using the most popular non-autoregressive TTS models; 4) FastSpeech 2s [Ren *et al.*, 2020]: the extension of FastSpeech 2 to fully end-to-end text-to-waveform generation with multi-task learning; 5) WaveGrad 2 [Chen *et al.*, 2021]: a diffusion probabilistic model to generate waveforms via gradient estimation.

The results are shown in Table 4: FastDiff-TTS could surpass competing end-to-end speech synthesis models and match the voice quality of the state-of-the-art cascaded TTS

Model	MOS
GT	4.52±0.09
GT(voc.)	4.28±0.07
Cascaded	4.13±0.07
FastSpeech 2s	3.94±0.06
WaveGrad 2	3.68±0.09
FastDiff-TTS	4.03±0.09

Table 4: Comparison with other text-to-speech models in quality.

systems, demonstrating that FastDiff-TTS is efficient in simplifying the overall text-to-speech synthesis pipeline.

6 Conclusion

In this work, we proposed FastDiff, a fast conditional diffusion model for high-quality speech synthesis. FastDiff employed a stack of time-aware location-variable convolutions with diverse receptive field patterns to model long-term time dependencies with adaptive conditions. A noise predictor was further adopted to derive tighter schedules for reducing reverse iterations without distinct quality degradation. The extension model FastDiff-TTS discarded intermediate features (e.g., spectrograms) and simplified the end-to-end text-to-waveform syntheses pipeline. Experimental results demonstrated that our proposed model outperformed the best publicly available models in terms of synthesis quality, even comparable to the human level. Moreover, FastDiff showed a significant improvement in synthesis speed, which required as few as 4 iterations to generate high-quality samples. To the best of our knowledge, FastDiff made diffusion models for the first time applicable to interactive, real-world speech generation with a low computational cost. In addition, FastDiff performed strong robustness and enjoyed high-quality synthesis in out-of-domain generalization to unseen speakers. We envisage that our work could serve as a basis for future speech synthesis studies.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2020YFC0832505, National Natural Science Foundation of China under Grant

No.61836002, No.62072397, Zhejiang Natural Science Foundation under Grant LR19F020006, information Technology Center and State Key Lab of CAD & CG, Zhejiang University, and Tencent AI Lab.

References

- [Chen *et al.*, 2020] 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.
- [Chen *et al.*, 2021] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*, 2021.
- [Creswell *et al.*, 2018] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 2018.
- [Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, volume 34, 2021.
- [Donahue *et al.*, 2020] Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech. In *Proc. of ICLR*, 2020.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.
- [Huang *et al.*, 2021] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proc. of ACM MM*, 2021.
- [Ito and Johnson, 2017] Keith Ito and Linda Johnson. The lj speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017. Accessed: 2022-01-01.
- [Jang *et al.*, 2021] 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.
- [Kim *et al.*, 2020] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Proc. of NeurIPS*, volume 33, pages 8067–8077, 2020.
- [Kim *et al.*, 2021] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. of ICML*, pages 5530–5540. PMLR, 2021.
- [Kong *et al.*, 2020a] 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, 2020.
- [Kong *et al.*, 2020b] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proc. of ICLR*, 2020.
- [Lam *et al.*, 2022] 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.
- [Liu *et al.*, 2021] 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.
- [Oord *et al.*, 2016] 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.
- [Prenger *et al.*, 2019] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. of ICASSP*, 2019.
- [Ren *et al.*, 2020] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *Proc. of ICLR*, 2020.
- [Ren *et al.*, 2021] Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech. In *Proc. of NeurIPS*, volume 34, 2021.
- [Rix *et al.*, 2001] 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.
- [Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. of ICLR*, 2020.
- [Taal *et al.*, 2010] 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.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017.
- [Xiao *et al.*, 2021] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.
- [Yamagishi *et al.*, 2019] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.
- [Zeng *et al.*, 2021] Zhen Zeng, Jianzong Wang, Ning Cheng, and Jing Xiao. Lvcnet: Efficient condition-dependent modeling network for waveform generation. In *Proc. of ICASSP*, 2021.