

DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism

Jinglin Liu*, Chengxi Li*, Yi Ren*, Feiyang Chen, Zhou Zhao[†]

Zhejiang University

{jinglinliu,chengxili,rayeren,zhaozhou}@zju.edu.cn, chenfeiyangai@gmail.com

Abstract

Singing voice synthesis (SVS) systems are built to synthesize high-quality and expressive singing voice, in which the acoustic model generates the acoustic features (*e.g.*, mel-spectrogram) given a music score. Previous singing acoustic models adopt a simple loss (*e.g.*, L1 and L2) or generative adversarial network (GAN) to reconstruct the acoustic features, while they suffer from over-smoothing and unstable training issues respectively, which hinder the naturalness of synthesized singing. In this work, we propose DiffSinger, an acoustic model for SVS based on the diffusion probabilistic model. DiffSinger is a parameterized Markov chain that iteratively converts the noise into mel-spectrogram conditioned on the music score. By implicitly optimizing variational bound, DiffSinger can be stably trained and generate realistic outputs. To further improve the voice quality and speed up inference, we introduce a shallow diffusion mechanism to make better use of the prior knowledge learned by the simple loss. Specifically, DiffSinger starts generation at a shallow step smaller than the total number of diffusion steps, according to the intersection of the diffusion trajectories of the ground-truth mel-spectrogram and the one predicted by a simple mel-spectrogram decoder. Besides, we propose boundary prediction methods to locate the intersection and determine the shallow step adaptively. The evaluations conducted on a Chinese singing dataset demonstrate that DiffSinger outperforms state-of-the-art SVS work. Extensional experiments also prove the generalization of our methods on text-to-speech task (DiffSpeech). Audio samples: <https://diffsinger.github.io>. Codes: <https://github.com/MoonInTheRiver/DiffSinger>.

1 Introduction

Singing voice synthesis (SVS) which aims to synthesize natural and expressive singing voice from musical score (Wu and Luan 2020), increasingly draws attention from the research community and entertainment industries (Zhang et al. 2020). The pipeline of SVS usually consists of an acoustic model to generate the acoustic features (*e.g.*, mel-spectrogram) conditioned on a music score, and a vocoder to convert the acoustic features to waveform (Nakamura et al.

2019; Lee et al. 2019; Blaauw and Bonada 2020; Ren et al. 2020; Chen et al. 2020)¹.

Previous singing acoustic models mainly utilize simple loss (*e.g.*, L1 or L2) to reconstruct the acoustic features. However, this optimization is based on the incorrect unimodal distribution assumptions, leading to blurry and over-smoothing outputs. Although existing methods endeavor to solve this problem by generative adversarial network (GAN) (Lee et al. 2019; Chen et al. 2020), training an effective GAN may occasionally fail due to the unstable discriminator. These issues hinder the naturalness of synthesized singing.

Recently, a highly flexible and tractable generative model, diffusion probabilistic model (a.k.a. diffusion model) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021) emerges. Diffusion model consists of two processes: diffusion process and reverse process (also called denoising process). The diffusion process is a Markov chain with fixed parameters (when using the certain parameterization in (Ho, Jain, and Abbeel 2020)), which converts the complicated data into isotropic Gaussian distribution by adding the Gaussian noise gradually; while the reverse process is a Markov chain implemented by a neural network, which learns to restore the origin data from Gaussian white noise iteratively. Diffusion model can be stably trained by implicitly optimizing variational lower bound (ELBO) on the data likelihood. It has been demonstrated that diffusion model can produce promising results in image generation (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021) and neural vocoder (Chen et al. 2021; Kong et al. 2021) fields.

In this work, we propose DiffSinger, an acoustic model for SVS based on diffusion model, which converts the noise into mel-spectrogram conditioned on the music score. DiffSinger can be efficiently trained by optimizing ELBO, without adversarial feedback, and generates realistic mel-spectrograms strongly matching the ground truth distribution.

To further improve the voice quality and speed up inference, we introduce a shallow diffusion mechanism to make better use of the prior knowledge learned by the simple loss. Specifically, we find that there is an intersection of the diffu-

*Equal contribution.

[†]Corresponding author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹A music score consists of lyrics, pitch and duration.

sion trajectories of the ground-truth mel-spectrogram M and the one predicted by a simple mel-spectrogram decoder \tilde{M} ²: sending M and \tilde{M} into the diffusion process could result in similar distorted mel-spectrograms, when the diffusion step is big enough (but not reaches the deep step where the distorted mel-spectrograms become Gaussian white noise). Thus, in the inference stage we 1) leverage the simple mel-spectrogram decoder to generate \tilde{M} ; 2) calculate the sample at a shallow step k through the diffusion process: \tilde{M}_k ³; and 3) start reverse process from \tilde{M}_k rather than Gaussian white noise, and complete the process by k iteration denoising steps (Vincent 2011; Song and Ermon 2019; Ho, Jain, and Abbeel 2020). Besides, we train a boundary prediction network to locate this intersection and determine the k adaptively. The shallow diffusion mechanism provides a better start point than Gaussian white noise and alleviates the burden of the reverse process, which improves the quality of synthesized audio and accelerates inference.

Finally, since the pipeline of SVS resembles that of text-to-speech (TTS) task, we also build DiffSpeech adjusting from DiffSinger for generalization. The evaluations conducted on a Chinese singing dataset demonstrate the superiority of DiffSinger (0.11 MOS gains compared with a state-of-the-art acoustic model for SVS (Wu and Luan 2020)), and the effectiveness of our novel mechanism (0.14 MOS gains, 0.5 CMOS gains and 45.1% speedup with shallow diffusion mechanism). The extensional experiments of DiffSpeech on TTS task prove the generalization of our methods (0.24/0.23 MOS gains compared with FastSpeech 2 (Ren et al. 2021) and Glow-TTS (Kim et al. 2020) respectively). The contributions of this work can be summarized as follows:

- We propose DiffSinger, which is the first acoustic model for SVS based on diffusion probabilistic model. DiffSinger addresses the over-smoothing and unstable training issues in previous works.
- We propose a shallow diffusion mechanism to further improve the voice quality, and accelerate the inference.
- The extensional experiments on TTS task (DiffSpeech) prove the generalization of our methods.

2 Diffusion Model

In this section, we introduce the theory of diffusion probabilistic model (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020). The full proof can be found in previous works (Ho, Jain, and Abbeel 2020; Kong et al. 2021; Song, Meng, and Ermon 2021). A diffusion probabilistic model converts the raw data into Gaussian distribution gradually by a diffusion process, and then learns the reverse process to restore the data from Gaussian white noise (Sohl-Dickstein et al. 2015). These processes are shown in Figure 1.

²Here we use a traditional acoustic model based on feed-forward Transformer (Ren et al. 2021; Blaauw and Bonada 2020), which is trained by L1 loss to reconstruct mel-spectrogram.

³ $K < T$, where T is the total number of diffusion steps. \tilde{M}_k can be calculated in closed form time (Ho, Jain, and Abbeel 2020).

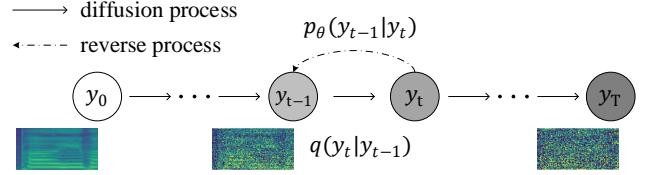


Figure 1: The directed graph for diffusion model.

Diffusion Process Define the data distribution as $q(\mathbf{y}_0)$, and sample $\mathbf{y}_0 \sim q(\mathbf{y}_0)$. The diffusion process is a Markov chain with fixed parameters (Ho, Jain, and Abbeel 2020), which converts \mathbf{y}_0 into the latent \mathbf{y}_T in T steps:

$$q(\mathbf{y}_{1:T}|\mathbf{y}_0) := \prod_{t=1}^T q(\mathbf{y}_T|\mathbf{y}_{t-1}).$$

At each diffusion step $t \in [1, T]$, a tiny Gaussian noise is added to \mathbf{y}_{t-1} to obtain \mathbf{y}_t , according to a variance schedule $\beta = \{\beta_1, \dots, \beta_T\}$:

$$q(\mathbf{y}_t|\mathbf{y}_{t-1}) := \mathcal{N}(\mathbf{y}_t; \sqrt{1 - \beta_t} \mathbf{y}_{t-1}, \beta_t \mathbf{I}).$$

If β is well designed and T is sufficiently large, then $q(\mathbf{y}_T)$ is nearly an isotropic Gaussian distribution (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). Besides, there is a special property of diffusion process that $q(\mathbf{y}_T|\mathbf{y}_0)$ can be calculated in closed form in $O(1)$ time (Ho, Jain, and Abbeel 2020):

$$q(\mathbf{y}_T|\mathbf{y}_0) = \mathcal{N}(\mathbf{y}_T; \sqrt{\bar{\alpha}_t} \mathbf{y}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (1)$$

where $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$, $\alpha_t := 1 - \beta_t$.

Reverse Process The reverse process is a Markov chain with learnable parameters θ from \mathbf{y}_T to \mathbf{y}_0 . Since the exact reverse transition distribution $q(\mathbf{y}_{t-1}|\mathbf{y}_t)$ is intractable, we approximate it by a neural network with parameters θ (θ is shared at every t -th step):

$$p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t) := \mathcal{N}(\mathbf{y}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{y}_t, t), \sigma_t^2 \mathbf{I}). \quad (2)$$

Thus the whole reverse process can be defined as:

$$p_\theta(\mathbf{y}_{0:T}) := p(\mathbf{y}_T) \prod_{t=1}^T p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t).$$

Training To learn the parameters θ , we minimize a variational bound of the negative log likelihood:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{y}_0)}[-\log p_\theta(\mathbf{y}_0)] &\geq \\ \mathbb{E}_{q(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_T)}[\log q(\mathbf{y}_{1:T}|\mathbf{y}_0) - \log p_\theta(\mathbf{y}_{0:T})] &=: \mathbb{L}. \end{aligned}$$

Efficient training is optimizing a random term of \mathbb{L} with stochastic gradient descent (Ho, Jain, and Abbeel 2020):

$$\mathbb{L}_{t-1} = D_{\text{KL}}(q(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{y}_0) \parallel p_\theta(\mathbf{y}_{t-1}|\mathbf{y}_t)), \quad (3)$$

where

$$\begin{aligned} q(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{y}_0) &= \mathcal{N}(\mathbf{y}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{y}_t, \mathbf{y}_0), \tilde{\beta}_t \mathbf{I}) \\ \tilde{\boldsymbol{\mu}}_t(\mathbf{y}_t, \mathbf{y}_0) &:= \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \mathbf{y}_0 + \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{y}_t, \end{aligned}$$

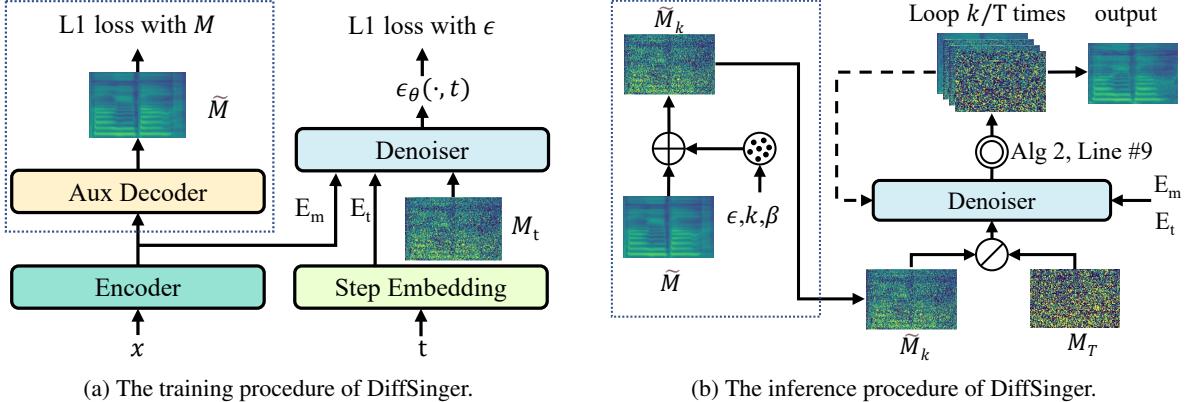


Figure 2: The overview of DiffSinger (with shallow diffusion mechanism in the dotted line boxes). In subfigure (a), x is the music score; t is the step number; M means the ground truth mel-spectrogram; \tilde{M} means the blurry mel-spectrogram generated by the auxiliary decoder trained with L1 loss; M_t is M at the t -th step in the diffusion process. In subfigure (b), M_T means the M at T -th diffusion step (Gaussian white noise); k is the predicted intersection boundary; there is a switch to select M_T (naive version) or \tilde{M}_k (with shallow diffusion) as the start point of the inference procedure.

where $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$. Eq. (3) is equivalent to:

$$\mathbb{L}_{t-1} - \mathcal{C} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(\mathbf{y}_t, \mathbf{y}_0) - \mu_\theta(\mathbf{y}_t, t)\|^2 \right], \quad (4)$$

where \mathcal{C} is a constant. And by reparameterizing Eq. (1) as $\mathbf{y}_t(\mathbf{y}_0, \epsilon) = \sqrt{\bar{\alpha}_t} \mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t} \epsilon$, and choosing the parameterization:

$$\mu_\theta(\mathbf{y}_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{y}_t, t) \right), \quad (5)$$

Eq. (4) can be simplified to:

$$\mathbb{E}_{\mathbf{y}_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t} \epsilon, t)\|^2 \right]. \quad (6)$$

Finally we set σ_t^2 to $\tilde{\beta}_t$, sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon_\theta(\cdot)$ is the outputs of the neural network.

Sampling Sample \mathbf{y}_T from $p(\mathbf{y}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and run the reverse process to obtain a data sample.

3 DiffSinger

As illustrated in Figure 2, DiffSinger is built on the diffusion model. Since SVS task models the conditional distribution $p_\theta(M_0|x)$, where M is the mel-spectrogram and x is the music score corresponding to M , we add x to the diffusion denoiser as the condition in the reverse process. In this section, we first describe a naive version of DiffSinger (Section 3.1); then we introduce a novel shallow diffusion mechanism to improve the model performance and efficiency (Section 3.2); finally, we describe the boundary prediction network which can adaptively find the intersection boundary required in shallow diffusion mechanism (Section 3.3).

3.1 Naive Version of DiffSinger

In the naive version of DiffSinger (without dotted line boxes in Figure 2): In the training procedure (shown in Figure 2a), DiffSinger takes in the mel-spectrogram at t -th step M_t in the diffusion process and predicts the random noise $\epsilon_\theta(\cdot)$ in Eq. (6), conditioned on t and the music score x . The inference procedure (shown in Figure 2b) starts at the Gaussian white noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, as the previous diffusion models do (Ho, Jain, and Abbeel 2020; Kong et al. 2021). Then the procedure iterates for T times to repeatedly denoise the intermediate samples with two steps: 1) predict the $\epsilon_\theta(\cdot)$ using the denoiser; 2) obtain M_{t-1} from M_t using the predicted $\epsilon_\theta(\cdot)$, according to Eq. (2) and Eq. (5):

$$M_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(M_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(M_t, x, t) \right) + \sigma_t \mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ when $t > 1$, and $\mathbf{z} = 0$ when $t = 1$. Finally, a mel-spectrogram M corresponding to x could be generated.

3.2 Shallow Diffusion Mechanism

Although the previous acoustic model trained by the simple loss has intractable drawbacks, it still generates samples showing strong connection⁴ to the ground-truth data distribution, which could provide plenty of prior knowledge to DiffSinger. To explore this connection and find a way to make better use of the prior knowledge, we conduct the empirical observation leveraging the diffusion process (shown in Figure 3): 1) when $t = 0$, M has rich details between the neighboring harmonics, which can influence the naturalness of the synthesized singing voice, but \tilde{M} is over-smoothing as we introduced in Section 1; 2) as t increases, samples

⁴The samples fail to maintain the variable aperiodic parameters, but they usually have a clear “skeleton” (harmonics) matching the ground truth.

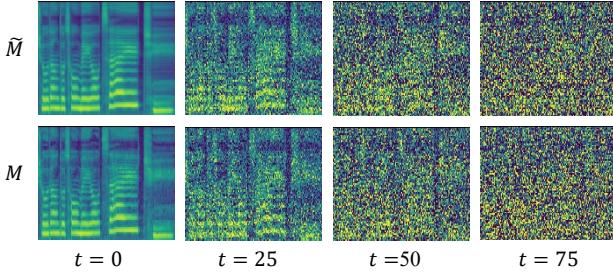


Figure 3: The mel-spectrograms at different steps in the diffusion process. The first line shows the diffusion process of mel-spectrograms \tilde{M} generated by a simple decoder trained with L1 loss; the second line shows that of ground truth mel-spectrograms.

of two process become indistinguishable. We illustrate this observation in Figure 4: the trajectory from \tilde{M} manifold to Gaussian noise manifold and the trajectory from M to Gaussian noise manifold intersect when the diffusion step is big enough.

Inspired by this observation, we propose the shallow diffusion mechanism: instead of starting with the Gaussian white noise, the reverse process starts at the intersection of two trajectories shown in Figure 4. Thus the burden of the reverse process could be distinctly alleviated⁵. Specifically, in the inference stage we 1) leverage an auxiliary decoder to generate \tilde{M} , which is trained with L1 conditioned on the music score encoder outputs, as shown in the dotted line box in Figure 2a; 2) generate the intermediate sample at a shallow step k through the diffusion process, as shown in the dotted line box in Figure 2b according to Eq. (1):

$$\tilde{M}_k(\tilde{M}, \epsilon) = \sqrt{\bar{\alpha}_k} \tilde{M} + \sqrt{1 - \bar{\alpha}_k} \epsilon,$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{\alpha}_k := \prod_{s=1}^k \alpha_s$, $\alpha_k := 1 - \beta_k$. If the intersection boundary k is properly chosen, it can be considered that \tilde{M}_k and M_k come from the same distribution; 3) start reverse process from \tilde{M}_k , and complete the process by k iterations denoising. The training and inference procedures with shallow diffusion mechanism are described in Algorithm 1 and 2 respectively. The theoretical proof of the intersection of two trajectories can be found in the supplement.

3.3 Boundary Prediction

We propose a boundary predictor (BP) to locate the intersection in Figure 4 and determine k adaptively. Concretely, BP consists of a classifier and a module for adding noise to mel-spectrograms according to Eq. (1). Given the step number $t \in [0, T]$, we label the M_t as 1 and \tilde{M}_t as 0, and use cross-entropy loss to train the boundary predictor to judge whether the input mel-spectrogram at t step in diffusion process comes from M or \tilde{M} . The training loss \mathbb{L}_{BP} can be

⁵Converting M_k into M_0 is easier than converting M_T (Gaussian white noise) into M_0 ($k < T$). Thus the former could improve the quality of synthesized audio and accelerates inference.

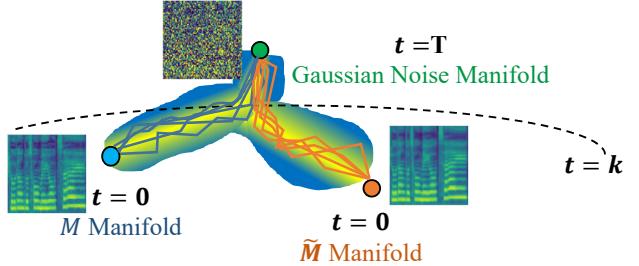


Figure 4: The diffusion trajectories of M and \tilde{M} . Two distributions $q(M_t|M_0)$ and $q(\tilde{M}_t|\tilde{M}_0)$ become closer as t increases.

written as:

$$\mathbb{L}_{BP} = -\mathbb{E}_{M \in \mathcal{Y}, t \in [0, T]} [\log BP(M_t, t) + \log(1 - BP(\tilde{M}_t, t))],$$

where \mathcal{Y} is the training set of mel-spectrograms. When BP have been trained, we determine k using the predicted value of BP, which indicates the probability of a sample classified to be 1. For all $M \in \mathcal{Y}$, we find the earliest step k' where the 95% steps t in $[k', T]$ satisfies: the margin between $BP(M_t, t)$ and $BP(\tilde{M}_t, t)$ is under the threshold. Then we choose the average of k' as the intersection boundary k .

Algorithm 1: Training procedure of DiffSinger.

Input: The denoiser ϵ_θ ; the intersection boundary k ; the training set $(\mathcal{X}, \mathcal{Y})$.

```

1 repeat
2   | Sample  $(x, M)$  from  $(\mathcal{X}, \mathcal{Y})$ ;
3   |  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ;
4   |  $t \sim \text{Uniform}(\{1, \dots, k\})$ ;
5   | Take gradient descent step on
6   |  $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} M + \sqrt{1 - \bar{\alpha}_t} \epsilon, x, t)\|^2$ 
7 until convergence;

```

We also propose an easier trick for boundary prediction in the supplement by comparing the KL-divergence. Note that the boundary prediction can be considered as a step of dataset preprocessing to choose the hyperparameter k for the whole dataset. k actually can be chosen manually by brute-force searching on validation set.

3.4 Model Structures

Encoder The encoder encodes the music score into the condition sequence, which consists of 1) a lyrics encoder to map the phoneme ID into embedding sequence, and a series of Transformer blocks (Vaswani et al. 2017) to convert this sequence into linguistic hidden sequence; 2) a length regulator to expand the linguistic hidden sequence to the length of mel-spectrograms according to the duration information; and 3) a pitch encoder to map the pitch ID into pitch embedding sequence. Finally, the encoder adds linguistic sequence and pitch sequence together as the music condition sequence E_m following (Ren et al. 2020).

Algorithm 2: Inference procedure of DiffSinger.

Input: The denoiser ϵ_θ ; the auxiliary decoder; the intersection boundary k ; the source testing set \mathcal{X} .

- 1 Sample x from \mathcal{X} as condition;
- 2 Generate \tilde{M} by the auxiliary decoder;
- 3 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
- 4 $\tilde{M}_k(\tilde{M}, \epsilon) = \sqrt{\bar{\alpha}_k} \tilde{M} + \sqrt{1 - \bar{\alpha}_k} \epsilon$;
- 5 $M_k = \tilde{M}_k$;
- 6 **for** $t = k, k - 1, \dots, 1$ **do**
- 7 **if** $t = 1$ **then** $\mathbf{z} = \mathbf{0}$;
- 8 **else** Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
- 9 $M_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(M_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(M_t, x, t) \right) + \sigma_t \mathbf{z}$
- 10 **end**

Step Embedding The diffusion step t is another conditional input for denoiser ϵ_θ , as shown in Eq. (6). To convert the discrete step t to continuous hidden, we use the sinusoidal position embedding (Vaswani et al. 2017) followed by two linear layers to obtain step embedding E_t with C channels.

Auxiliary Decoder We introduce a simple mel-spectrogram decoder called the auxiliary decoder, which is composed of stacked feed-forward Transformer (FFT) blocks and generates \tilde{M} as the final outputs, the same as the mel-spectrogram decoder in FastSpeech 2 (Ren et al. 2021).

Denoiser Denoiser ϵ_θ takes in M_t as input to predict ϵ added in diffusion process conditioned on the step embedding E_t and music condition sequence E_m . Since diffusion model imposes no architectural constraints (Sohl-Dickstein et al. 2015; Kong et al. 2021), the design of denoiser has multiple choices. We adopt a non-causal WaveNet (Oord et al. 2016) architecture proposed by (Rethage, Pons, and Serra 2018; Kong et al. 2021) as our denoiser. The denoiser is composed of a 1×1 convolution layer to project M_t with H_m channels to the input hidden sequence \mathcal{H} with C channels and N convolution blocks with residual connections. Each convolution block consists of 1) an element-wise adding operation which adds E_t to \mathcal{H} ; 2) a non-causal convolution network which converts \mathcal{H} from C to $2C$ channels; 3) a 1×1 convolution layer which converts the E_m to $2C$ channels; 4) a gate unit to merge the information of input and conditions; and 5) a residual block to split the merged hidden into two branches with C channels (the residual as the following \mathcal{H} and the “skip hidden” to be collected as the final results), which enables the denoiser to incorporate features at several hierarchical levels for final prediction.

Boundary Predictor The classifier in the boundary predictor is composed of 1) a step embedding to provide E_t ; 2) a ResNet (He et al. 2016) with stacked convolutional layers and a linear layer, which takes in the mel-spectrograms at t -th step and E_t to classify M_t and \tilde{M}_t .

More details of model structure and configurations are

shown in the supplement.

4 Experiments

In this section, we first describe the experimental setup, and then provide the main results on SVS with analysis. Finally, we conduct the extensional experiments on TTS.

4.1 Experimental Setup

Dataset Since there is no publicly available high-quality unaccompanied singing dataset, we collect and annotate a Chinese Mandarin pop songs dataset: PopCS, to evaluate our methods. PopCS contains 117 Chinese pop songs (total ~ 5.89 hours with lyrics) collected from a qualified female vocalist. All the audio files are recorded in a recording studio. Every song is sampled at 24kHz with 16-bit quantization. To obtain more accurate music scores corresponding to the songs (Lee et al. 2019), we 1) split each whole song into sentence pieces following DeepSinger (Ren et al. 2020) and train a Montreal Forced Aligner tool (MFA) (McAuliffe et al. 2017) model on those sentence-level pairs to obtain the phoneme-level alignments between song piece and its corresponding lyrics; 2) extract F_0 (fundamental frequency) as pitch information from the raw waveform using Parselmouth, following (Wu and Luan 2020; Blaauw and Bonada 2020; Ren et al. 2020). We randomly choose 2 songs for validation and testing. To release a high-quality dataset, after the paper is accepted, we clean and re-segment these songs, resulting in 1,651 song pieces, which mostly last 10~13 seconds. The codes accompanied with the access to PopCS are in <https://github.com/MoonInTheRiver/DiffSinger>. In this repository, we also add the extra codes out of interest, for MIDI-to-Mel, including the MIDI-to-Mel without F_0 prediction/condition.

Implementation Details We convert Chinese lyrics into phonemes by pypinyin following (Ren et al. 2020); and extract the mel-spectrogram (Shen et al. 2018) from the raw waveform; and set the hop size and frame size to 128 and 512 in respect of the sample rate 24kHz. The size of phoneme vocabulary is 61. The number of mel bins H_m is 80. The mel-spectrograms are linearly scaled to the range $[-1, 1]$, and F_0 is normalized to have zero mean and unit variance. In the lyrics encoder, the dimension of phoneme embeddings is 256 and the Transformer blocks have the same setting as that in FastSpeech 2 (Ren et al. 2021). In the pitch encoder, the size of the lookup table and encoded pitch embedding are set to 300 and 256. The channel size C mentioned before is set to 256. In the denoiser, the number of convolution layers N is 20 with the kernel size 3, and we set the dilation to 1 (without dilation) at each layer⁶. We set T to 100 and β to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.06$. The auxiliary decoder has the same setting as the mel-spectrogram decoder in FastSpeech 2. In the boundary predictor, the number of convolutional layers is 5, and the threshold is set to 0.4 empirically.

⁶You can consider setting a bigger dilation number to increase the receptive field of the denoiser. See our Github repository.

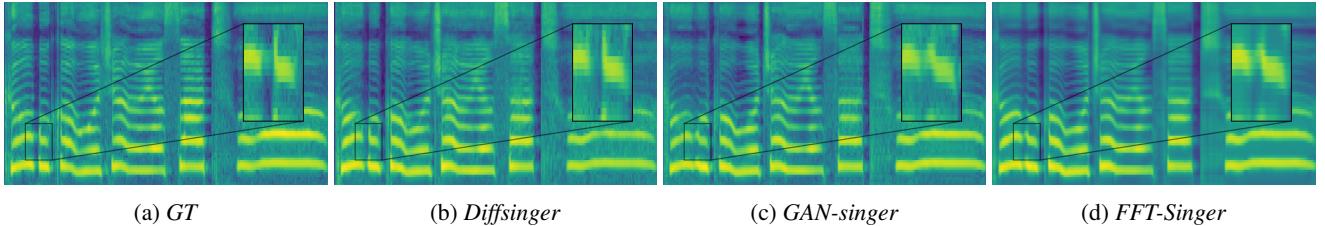


Figure 5: Visualizations of mel-spectrograms in four systems: GT, DiffSinger, GAN-Singer and FFT-Singer.

Training and Inference The training has two stages: 1) warmup stage: separately train the auxiliary decoder for 160k steps with the music score encoder, and then leverage the auxiliary decoder to train the boundary predictor for 30k steps to obtain k ; 2) main stage: training DiffSinger as Algorithm 1 describes for 160k steps until convergence. In the inference stage, for all SVS experiments, we uniformly use a pretrained Parallel WaveGAN (PWG) (Yamamoto, Song, and Kim 2020)⁷ as vocoder to transform the generated mel-spectrograms into waveforms (audio samples).

4.2 Main Results and Analysis

Audio Performance To evaluate the perceptual audio quality, we conduct the MOS (mean opinion score) evaluation on the test set. Eighteen qualified listeners are asked to make judgments about the synthesized song samples. We compare MOS of the song samples generated by DiffSinger with the following systems: 1) *GT*, the ground truth singing audio; 2) *GT (Mel + PWG)*, where we first convert the ground truth singing audio to the ground truth mel-spectrograms, and then convert these mel-spectrograms back to audio using PWG vocoder described in Section 4.1; 3) *FFT-NPSS (Blaauw and Bonada 2020) (WORLD)*, the SVS system which generates WORLD vocoder features (Morise, Yokomori, and Ozawa 2016) through feed-forward Transformer (FFT) and uses WORLD vocoder to synthesize audio; 4) *FFT-Singer (Mel + PWG)* the SVS system which generates mel-spectrograms through FFT network and uses PWG vocoder to synthesize audio; 5) *GAN-Singer (Wu and Luan 2020) (Mel + PWG)*, the SVS system with adversarial training using multiple random window discriminators.

The results are shown in Table 1. The quality of *GT (MEL + PWG)* (4.04 ± 0.11) is the upper limit of the acoustic model for SVS. *DiffSinger* outperforms the baseline system with simple training loss (*FFT-Singer*) by a large margin, and shows the superiority compared with the state-of-the-art GAN-based method (*GAN-Singer* (Wu and Luan 2020)), which demonstrate the effectiveness of our method.

As shown in Figure 5, we compare the ground truth, the generated mel-spectrograms from *DiffSinger*, *GAN-singer* and *FFT-Singer* with the same music score. It can be seen that both Figure 5c and Figure 5b contain more delicate details between harmonics than Figure 5d does. Moreover, the

⁷We adjust PWG to take in F0 driven source excitation (Wang and Yamagishi 2020) as additional condition, similar to that in (Chen et al. 2020).

Method	MOS
<i>GT</i>	4.30 ± 0.09
<i>GT (Mel + PWG)</i>	4.04 ± 0.11
<i>FFT-NPSS (WORLD)</i>	1.75 ± 0.17
<i>FFT-Singer (Mel + PWG)</i>	3.67 ± 0.11
<i>GAN-Singer (Mel + PWG)</i>	3.74 ± 0.12
<i>DiffSinger Naive (Mel + PWG)</i>	3.71 ± 0.10
<i>DiffSinger (Mel + PWG)</i>	3.85 ± 0.11

Table 1: The MOS with 95% confidence intervals of song samples. DiffSinger Naive means the naive version of DiffSinger without shallow diffusion mechanism.

performance of *DiffSinger* in the region of mid or low frequency is more competitive than that of *GAN-singer* while maintaining similar quality of the high-frequency region.

In the meanwhile, the shallow diffusion mechanism accelerates inference of naive diffusion model by 45.1% (RTF 0.191 vs. 0.348, RTF is the real-time factor, that is the seconds it takes to generate one second of audio).

Ablation Studies We conduct ablation studies to demonstrate the effectiveness of our proposed methods and some hyper-parameters studies to seek the best model configurations. We conduct CMOS evaluation for these experiments. The results of variations on *DiffSinger* are listed in Table 2. It can be seen that: 1) removing the shallow diffusion mechanism results in quality drop (-0.500 CMOS), which is consistent with the MOS test results and verifies the effectiveness of our shallow diffusion mechanism (row 1 vs. row 2); 2) adopting other k (row 1 vs. row 3) rather than the one predicted by our boundary predictor causes quality drop, which verifies that our boundary prediction network can predict a proper k for shallow diffusion mechanism; and 3) the model with configurations $C = 256$ and $L = 20$ produces the best results (row 1 vs. row 4,5,6,7), indicating that our model capacity is sufficient.

4.3 Extensional Experiments on TTS

To verify the generalization of our methods on TTS task, we conduct the extensional experiments on LJSpeech dataset (Ito and Johnson 2017), which contains 13,100 English audio clips (total ~ 24 hours) with corresponding transcripts. We follow the train-val-test dataset splits, the pre-processing of mel-spectrograms, and the grapheme-to-phoneme tool in FastSpeech 2. To build *DiffSpeech*, we 1)

No.	C	L	w/ shallow	k	CMOS
1	256	20	✓	54	0.000
2	256	20	✗	-	-0.500
3	256	20	✓	25	-0.053
4	128	20	✓	54	-0.071
5	512	20	✓	54	-0.044
6	256	10	✓	54	-0.293
7	256	30	✓	54	-0.445

Table 2: Variations on the *DiffSinger*. T in all the experiments is set to 100. C is channel size; L is the number of layers in denoiser; w/ shallow means “with shallow diffusion mechanism”; $k = 54$ is our predicted intersection boundary.

Method	MOS
<i>GT</i>	4.22 ± 0.07
<i>GT</i> (<i>Mel + HiFi-GAN</i>)	4.15 ± 0.07
<i>Tacotron 2</i> (<i>Mel + HiFi-GAN</i>)	3.54 ± 0.05
<i>BVAE-TTS</i> (<i>Mel + HiFi-GAN</i>)	3.48 ± 0.06
<i>FastSpeech 2</i> (<i>Mel + HiFi-GAN</i>)	3.68 ± 0.06
<i>Glow-TTS</i> (<i>Mel + HiFi-GAN</i>)	3.69 ± 0.07
<i>DiffSpeech Naive</i> (<i>Mel + HiFi-GAN</i>)	3.69 ± 0.05
<i>DiffSpeech</i> (<i>Mel + HiFi-GAN</i>)	3.92 ± 0.06

Table 3: The MOS of speech samples with 95% confidence intervals.

add a pitch predictor and a duration predictor to *DiffSinger* as those in FastSpeech 2; 2) adopt $k = 70$ for shallow diffusion mechanism.

We use Amazon Mechanical Turk (ten testers) to make subjective evaluation and the results are shown in Table 3. All the systems adopt HiFi-GAN (Kong, Kim, and Bae 2020) as vocoder. *DiffSpeech* outperforms FastSpeech 2 and Glow-TTS, which demonstrates the generalization. Besides, the last two rows in Table 3 also show the effectiveness of shallow diffusion mechanism (with 29.2% speedup, RTF 0.121 vs. 0.171).

5 Related Work

5.1 Singing Voice Synthesis

Initial works of singing voice synthesis generate the sounds using concatenated (Macon et al. 1997; Kenmochi and Ohshita 2007) or HMM-based parametric (Saino et al. 2006; Oura et al. 2010) methods, which are kind of cumbersome and lack flexibility and harmony. Thanks to the rapid evolution of deep learning, several SVS systems based on deep neural networks have been proposed in the past few years. Nishimura et al. (2016); Blaauw and Bonada (2017); Kim et al. (2018); Nakamura et al. (2019); Gu et al. (2020) utilize neural networks to map the contextual features to acoustic features. Ren et al. (2020) build the SVS system from scratch using singing data mined from music websites. Blaauw and Bonada (2020) propose a feed-forward Trans-

former SVS model for fast inference and avoiding exposure bias issues caused by autoregressive models. Besides, with the help of adversarial training, Lee et al. (2019) propose an end-to-end framework which directly generates linear-spectrograms. Wu and Luan (2020) present a multi-singer SVS system with limited available recordings and improve the voice quality by adding multiple random window discriminators. Chen et al. (2020) introduce multi-scale adversarial training to synthesize singing with a high sampling rate (48kHz). The voice naturalness and diversity of SVS system have been continuously improved in recent years.

5.2 Denoising Diffusion Probabilistic Models

A diffusion probabilistic model is a parameterized Markov chain trained by optimizing variational lower bound, which generates samples matching the data distribution in constant steps (Ho, Jain, and Abbeel 2020). Diffusion model is first proposed by Sohl-Dickstein et al. (2015). Ho, Jain, and Abbeel (2020) make progress of diffusion model to generate high-quality images using a certain parameterization and reveal an equivalence between diffusion model and denoising score matching (Song and Ermon 2019; Song et al. 2021). Recently, Kong et al. (2021) and Chen et al. (2021) apply the diffusion model to neural vocoders, which generate high-fidelity waveform conditioned on mel-spectrogram. Chen et al. (2021) also propose a continuous noise schedule to reduce the inference iterations while maintaining synthesis quality. Song, Meng, and Ermon (2021) extend diffusion model by providing a faster sampling mechanism, and a way to interpolate between samples meaningfully. Diffusion model is a fresh and developing technique, which has been applied in the fields of unconditional image generation, conditional spectrogram-to-waveform generation (neural vocoder). And in our work, we propose a diffusion model for the acoustic model which generates mel-spectrogram given music scores (or text). There is a concurrent work (Jeong et al. 2021) at the submission time of our preprint which adopts a diffusion model as the acoustic model for TTS task.

6 Conclusion

In this work, we proposed DiffSinger, an acoustic model for SVS based on diffusion probabilistic model. To improve the voice quality and speed up inference, we proposed a shallow diffusion mechanism. Specifically, we found that the diffusion trajectories of M and \tilde{M} converge together when the diffusion step is big enough. Inspired by this, we started the reverse process at the intersection (step k) of two trajectories rather than at the very deep diffusion step T . Thus the burden of the reverse process could be distinctly alleviated, which improves the quality of synthesized audio and accelerates inference. The experiments conducted on PopCS demonstrate the superiority of DiffSinger compared with previous works, and the effectiveness of our novel shallow diffusion mechanism. The extensional experiments conducted on LJSpeech dataset prove the effectiveness of DiffSpeech on TTS task. The directly synthesis without vocoder will be future work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2020YFC0832505, No.62072397, Zhejiang Natural Science Foundation under Grant LR19F020006. Thanks participants of the listening test for the valuable evaluations.

References

- Blaauw, M.; and Bonada, J. 2017. A neural parametric singing synthesizer modeling timbre and expression from natural songs. *Applied Sciences*, 7(12): 1313.
- Blaauw, M.; and Bonada, J. 2020. Sequence-to-sequence singing synthesis using the feed-forward transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7229–7233. IEEE.
- Chen, J.; Tan, X.; Luan, J.; Qin, T.; and Liu, T.-Y. 2020. HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis. *arXiv preprint arXiv:2009.01776*.
- Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; and Chan, W. 2021. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*.
- Gu, Y.; Yin, X.; Rao, Y.; Wan, Y.; Tang, B.; Zhang, Y.; Chen, J.; Wang, Y.; and Ma, Z. 2020. ByteSing: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoders. *arXiv preprint arXiv:2004.11012*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6832–6843. Curran Associates, Inc.
- Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. <https://keithito.com/LJ-Speech-Dataset/>. Accessed: 2019-10-12.
- Jeong, M.; Kim, H.; Cheon, S. J.; Choi, B. J.; and Kim, N. S. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*.
- Kenmochi, H.; and Ohshita, H. 2007. Vocaloid-commercial singing synthesizer based on sample concatenation. In *Eighth Annual Conference of the International Speech Communication Association*.
- Kim, J.; Choi, H.; Park, J.; Kim, S.; Kim, J.; and Hahn, M. 2018. Korean Singing Voice Synthesis System based on an LSTM Recurrent Neural Network. In *INTERSPEECH 2018*. ISCA.
- Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Advances in Neural Information Processing Systems*, 33.
- Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, 33.
- Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.
- Lee, J.; Choi, H.-S.; Jeon, C.-B.; Koo, J.; and Lee, K. 2019. Adversarially Trained End-to-End Korean Singing Voice Synthesis System. *Proc. Interspeech 2019*, 2588–2592.
- Macon, M.; Jensen-Link, L.; George, E. B.; Oliverio, J.; and Clements, M. 1997. Concatenation-based MIDI-to-singing voice synthesis. In *Audio Engineering Society Convention 103*. Audio Engineering Society.
- McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; and Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, 498–502.
- Morise, M.; Yokomori, F.; and Ozawa, K. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7): 1877–1884.
- Nakamura, K.; Hashimoto, K.; Oura, K.; Nankaku, Y.; and Tokuda, K. 2019. Singing voice synthesis based on convolutional neural networks. *arXiv preprint arXiv:1904.06868*.
- Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.
- Nishimura, M.; Hashimoto, K.; Oura, K.; Nankaku, Y.; and Tokuda, K. 2016. Singing Voice Synthesis Based on Deep Neural Networks. In *Interspeech*, 2478–2482.
- Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, 125–125.
- Oura, K.; Mase, A.; Yamada, T.; Muto, S.; Nankaku, Y.; and Tokuda, K. 2010. Recent development of the HMM-based singing voice synthesis system—Sinsky. In *Seventh ISCA Workshop on Speech Synthesis*.
- Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.
- Ren, Y.; Tan, X.; Qin, T.; Luan, J.; Zhao, Z.; and Liu, T.-Y. 2020. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1979–1989.
- Rethage, D.; Pons, J.; and Serra, X. 2018. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5069–5073. IEEE.
- Saino, K.; Zen, H.; Nankaku, Y.; Lee, A.; and Tokuda, K. 2006. An HMM-based singing voice synthesis system. In

Ninth International Conference on Spoken Language Processing.

- Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP 2018*, 4779–4783. IEEE.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, 2256–2265.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.
- Vincent, P. 2011. A Connection Between Score Matching and Denoising Autoencoders. In *Neural Computation*.
- Wang, X.; and Yamagishi, J. 2020. Using Cyclic Noise as the Source Signal for Neural Source-Filter-Based Speech Waveform Model. *Proc. Interspeech 2020*, 1992–1996.
- Wu, J.; and Luan, J. 2020. Adversarially Trained Multi-Singer Sequence-to-Sequence Singing Synthesizer. *Proc. Interspeech 2020*, 1296–1300.
- Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6199–6203. IEEE.
- Zhang, L.; Yu, C.; Lu, H.; Weng, C.; Zhang, C.; Wu, Y.; Xie, X.; Li, Z.; and Yu, D. 2020. DurIAN-SC: Duration Informed Attention Network Based Singing Voice Conversion System. *Proc. Interspeech 2020*, 1231–1235.

A Theoretical Proof of Intersection

Given a data sample M_0 and its corresponding \widetilde{M}_0 , the conditional distributions of M_t and \widetilde{M}_t are:

$$q(M_t|M_0) = \mathcal{N}(M_t; \sqrt{\bar{\alpha}_t}M_0, (1 - \bar{\alpha}_t)\mathbf{I})$$

$$q(\widetilde{M}_t|\widetilde{M}_0) = \mathcal{N}(\widetilde{M}_t; \sqrt{\bar{\alpha}_t}\widetilde{M}_0, (1 - \bar{\alpha}_t)\mathbf{I})$$

respectively. The KL-divergence between two Gaussian distributions is:

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}_0||\mathcal{N}_1) &= \frac{1}{2}[\text{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0) \\ &\quad - k + \ln(\frac{\det \Sigma_1}{\det \Sigma_0})] \end{aligned}$$

where k is the dimension; μ_0, μ_1 are means; Σ_0, Σ_1 are covariance matrices. Thus, in our case:

$$D_{\text{KL}}(\mathcal{N}(M_t)||\mathcal{N}(\widetilde{M}_t)) = \frac{\bar{\alpha}_t}{2(1 - \bar{\alpha}_t)} \|\widetilde{M}_0 - M_0\|_2^2$$

Since $\frac{\bar{\alpha}_t}{2(1 - \bar{\alpha}_t)}$ decreases towards 0 rapidly as t increases, this KL-divergence also decreases towards 0 rapidly as t increases. This guarantees the intersection of trajectories of the diffusion process.

Moreover, since the auxiliary decoder has been optimized by simple reconstruction loss (L1/L2 mentioned in the main paper) on the training set, $\|\widetilde{M}_0 - M_0\|_2^2$ is optimized towards minimum, which facilitates this intersection. In addition, \widetilde{M}_k does not need to be exactly the same as M_k , but just needs to come from vicinity of the mode of $q(M_k|M_0)$ (according to the theories of score matching and Langevin dynamics).

B An Easier Trick for Boundary Prediction

Intuitively, we can just adopt the smallest t as k when t satisfies:

$$\begin{aligned} \mathbb{E}_{M \in \mathcal{Y}'} [D_{\text{KL}}(\mathcal{N}(M_t)||\mathcal{N}(\widetilde{M}_t))] \\ = \mathbb{E}_{M \in \mathcal{Y}'} \left[\frac{\bar{\alpha}_t}{2(1 - \bar{\alpha}_t)} \|\widetilde{M}_0 - M_0\|_2^2 \right] \\ \leq \mathbb{E}_{M \in \mathcal{Y}'} [D_{\text{KL}}(\mathcal{N}(M_T)||\mathcal{N}(\mathbf{0}, \mathbf{I}))], \end{aligned}$$

which means that this start point at step k is at least not worse than the original prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. \mathcal{Y}' mean the mel-spectrograms in the validation set. In addition, when using this trick to determine k in DiffSpeech, it is more rational to generate \widetilde{M}_t ($\in \mathcal{Y}'$) conditioned on the ground-truth F0-contour & duration rather than the ones predicted.

C Details of Model Structure and Supplementary Configurations

C.1 Details of model structure

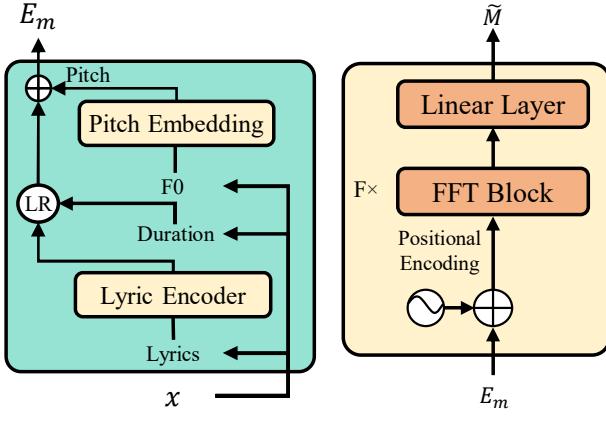
The detailed model structure of encoder, auxiliary decoder and denoiser are shown in Figure 6a, Figure 6b and Figure 7 respectively.

C.2 Supplementary configurations

In each FFT block: the number of FFT layers (F in Figure 6b) is set to 4; the hidden size of self-attention layer is 256; the number of attention heads is 2; the kernel sizes of 1D-convolution in the 2-layer convolutional layers are set to 9 and 1.

D Model Size

The model footprints of main systems for comparison in our paper are shown in Table 4. It can be seen that DiffSinger has the similar learnable parameters as other state-of-the-art models.



(a) Encoder.

(b) Auxiliary Decoder.

Figure 6: The detailed model structure of Encoder and Auxiliary Decoder. x is the music score. E_m is the music condition sequence. \tilde{M} means the blurry mel-spectrogram generated by the auxiliary decoder trained with L1 loss.

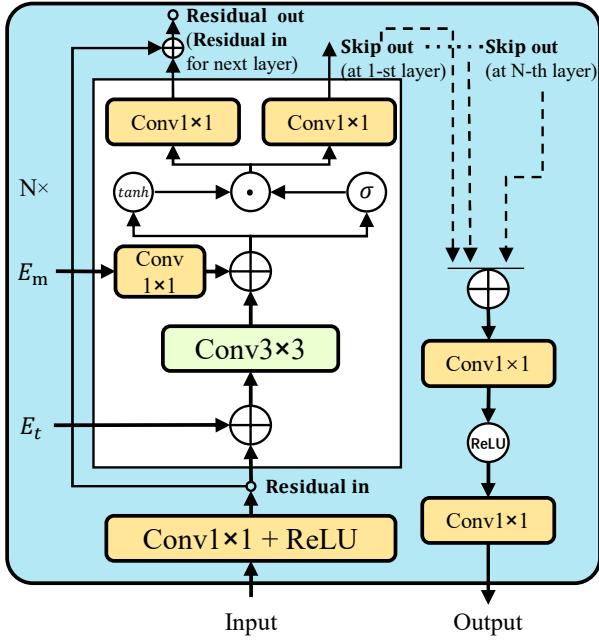


Figure 7: The detailed model structure of Denoiser. E_t is step embedding and E_m is music condition sequence. N is the number of residual layers. The model structure is derived from non-causal WaveNet, but simplified by replacing dilation layer to naive convolution layer.

E Details of Training and Inference

We train DiffSinger on 1 NVIDIA V100 GPU with 48 batch size. We adopt the Adam optimizer with learning rate $lr = 10^{-3}$. During training, the warmup stage costs about 16 hours and the main stage costs about 12 hours; During inference, the RTF of acoustic model for SVS and TTS are 0.191 and 0.121 respectively.

Model	Param(M)
<i>SVS Models</i>	
DiffSinger	26.744
FFT-Singer	24.254
GAN-Singer	24.254 (Generator) 0.963 (Discriminator)
<i>TTS Models</i>	
DiffSinger	27.722
Tacotron 2	28.193
BVAE-TTS	15.991
FastSpeech 2	24.179
Glow-TTS	28.589

Table 4: The model footprints. Param means the learnable parameters.