Investigation on Shallow Diffusion Mechanism in Singing Voice Synthesis

Hongjian Yu
hjyu@uw.edu

Ray Song
syx1995@uw.edu

Ziqing (Aurora) Yin
zyin5@uw.edu

1. Introduction

Singing Voice Synthesis (SVS) is a speech-processing task that benefits from the evolving scene of natural language and image generation. Past approaches to the task oftentimes apply sequence-to-sequence Generative Adversarial Network (GAN) models to generate spectrograms as images [6]. The task requires an acoustic model to generate key features, which are embedded by visual representations such as a mel-spectrogram [1], then feed the representations into a vocoder. However, earlier models were seldom considered to be competent in the task as they were not equipped with reasonable regularization. Unnatural and harsh sounds due to over-smoothing were common in the generated singing segments [4].

DiffSinger [4] is among the first few attempts to apply diffusion probabilistic models on audio and speech processing. In training cycles, it diffuses noise onto mel-spectrograms to motivate the generator to optimize the variational lower bound (ELBO) without the need for adversarial feedback. DiffSinger's approach not only generates more realistic mel-spectrograms but also introduces a shallow diffusion mechanism to improve the outputs and reduce training and inference time.

However, the intricate architecture of the model, especially the nuances of the diffusion and shallow diffusion mechanisms, poses substantial difficulties in interpretation. These complexities hinder our ability to thoroughly visualize and analyze specific model features, thereby impeding our progress in deriving valuable insights from the model's behavior and performance.

In this project, we want to investigate the various implementation choices within DiffSinger model along with systematic evaluation methodologies. We will adopt the training dataset and the general architecture presented in the paper. With a full replication of the original DiffSinger model, we plan to commit a few modifications to the boundary prediction mechanism, which is now handled by a neural network to identify the intersection of diffusion and reverse trajectories. By proposing a new method for computing the number of steps in each diffusion process, we are able to compare and contrast DiffSinger models with different settings, and consequently demonstrate the effect of introducing shallowness into the diffusion model.

While the original paper evaluated the performance of the model with a heavy reliance on subjective human assessments which poses limitations in objectivity and scalability, we aim to devise an evaluation strategy that encompasses both quantitative measures and qualitative judgments. Our experimental framework is designed to cohesively test the impact of varying diffusion steps on the DiffSinger model, aiming to refine our understanding and application of the diffusion mechanism in singing voice synthesis.

We will conduct a series of experiments with three distinct configurations of the model, each trained with a different number of diffusion steps. The configurations will be labeled as follows: Model A (low number of steps), Model B (medium number of steps), and Model C (high number of steps). This approach allows us to examine the trade-offs between synthesis quality and computational efficiency across different model complexities.

We will also design an evaluation scheme to bridge the gap between objective measurements and subjective qualities that contribute to a convincing singing voice. This method encompasses a suite of quantitative metrics aimed at precisely measuring the impact on synthesis quality and computational efficiency:

- Pitch Perturbation: Measurement of how accurately the model reproduces the pitch perturbations found in human singers.
- Timbre: Evaluation of the richness of synthesized voice texture.
- Dynamics: Expression range of the synthesized vocal.
- Robustness: Examination of the model's ability to generate extremely high and low articulations that are not present in the training data without artifacts.
- Noise: Quantification of any extraneous noise introduced during the synthesis process, which could detract from the naturalness of the singing voice.

Each model variant will undergo a training regimen tailored to its specific configuration. We will document the training duration, data requirements, and computational resources to evaluate the scalability and practicality of each configuration. Special attention will be paid to the convergence behavior of each model, noting any difficulties or peculiarities in the training process.

To complement our quantitative analyses, we will conduct human evaluation studies to assess the perceptual quality of the synthesized singing. Participants with varied musical backgrounds will be asked to rate the singing voices on naturalness, emotional expression, and overall listening experience. These subjective assessments will provide critical insights into the models' performance from the listener's perspective [4].

Through this research, we aim to deliver two key outcomes: firstly, an in-depth comparative analysis between shallow and standard diffusion models to discern their efficacy and performance in singing voice synthesis. Secondly, we intend to establish a robust, objective benchmark for evaluating SVS models, moving beyond subjective human judgment to ensure more reliable and consistent assessment criteria. This endeavor will provide valuable insights into optimizing the diffusion process for enhanced quality and computational efficiency in singing voice synthesis.

2. Methodology

2.1. Dataset

In the context of this research project, we employed the publicly available dataset: Opencpop. This dataset comprises a collection of 100 unique Mandarin Chinese pop songs, all performed by a professional female vocalist. The total length of the audio recordings amounts to approximately 5.2 hours, recorded in a standard recording studio at a 44,100 Hz sampling rate. The Opencpop dataset includes both MIDI and TextGrid annotations tailored for singing voice synthesis tasks, enhancing its utility for our model's training and evaluation. To improve efficiency, the dataset has been segmented into 5-second fragments, streamlining the processing and analysis phases of our work.

2.2. Diffusion

We investigate the integration of the diffusion probabilistic model into the SVS system, termed Diffsinger (Liu et al. 2022) [4]. Following the framework proposed by Ho, Jain, and Abbeel (2020) [2], the model iteratively adds and removes Gaussian noise over T steps, transitioning between the data and a latent Gaussian distribution. We adopt a variance schedule β to control the noise levels, with the process computationally optimized to allow efficient synthesis. More concrete diffusion steps can be found in previous work (Liu et al. 2022 [4]; Ho, Jain, and Abbeel 2020 [2]). This process ensures that, with appropriate β and large T, the resulting distribution of y_T can be approximated with a Gaussian distribution.

$$q(y_t|y_0) = \mathcal{N}(y_t; \sqrt{\overline{\alpha}_t} y_0, (1-\overline{\alpha}_t)I), \tag{1}$$

where $\overline{\alpha}$ is the cumulative product of $1-\beta$ up to time t.

The reverse denoising process described in the Diffsinger [4] paper is a Markov chain that transitions from the latent variable y_T back to the data y_0 using learnable parameter θ. The approximation at each step is expressed as a Gaussian distribution with mean $\mu_0(y_t, t)$ and variance $\sigma_t^2 I$. For the individual step transition:

$$p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, t), \sigma_t^2 I). \tag{2}$$

For the complete reverse process:

$$p_\theta(y_0 : y_T) = p(y_T) \prod_{t=1}^{T} p_\theta(y_{t-1}|y_t). \tag{3}$$

2.3. Shallow Diffusion

In our study, we adopted the shallow diffusion mechanism proposed in the DiffSinger framework (Liu et al. 2022) [4]. This term refers to a strategy employed to enhance the synthesis of the singing voice by leveraging the strengths of a pre-existing basic decoder model. The basic decoder trained with a simple loss function produces outputs that share significant structural similarities with the ground-truth data distribution. However, the model tends to over-smooth the outputs, denoted as $\tilde{M}$, leading to a loss of detail. Upon observing the diffusion process of both the decoder outputs $\tilde{M}$ and the ground-truth M, it was noted that as the diffusion steps increase, the differences between the two sets of outputs diminish. At a sufficiently advanced step, the outputs from the two processes become indistinguishable. This intersection significantly reduces the complexity of the reverse process. During inference, an auxiliary decoder generates the initial simplified output $\tilde{M}$, conditioned on the music score encoder's outputs. An intermediate sample is then produced at a shallow diffusion step k, using the relationship:

$$\tilde{M}_k(M, \epsilon) = \sqrt{\overline{\alpha}_k}\tilde{M} + \sqrt{1-\overline{\alpha}_k}\epsilon, \tag{4}$$

where ϵ is drawn from a normal distribution, and $\overline{\alpha_k}$is the product of the noise coefficients up to step k.

By generating an intermediate sample at a shallow step and proceeding with the reverse process from there, the approach leverages the structure within the simplified outputs to more efficiently synthesize the singing voice.

The shallow diffusion approach presupposes the simplified outputs ($\tilde{M}$) and the true data outputs (M) from the basic acoustic model approximate the same distribution at a certain diffusion step k. This approximation is critical as it predicates the starting point for the reverse diffusion process, which aims to reconstruct the original signal with reduced computational effort. The full rationale and proof of the trajectories intersection, which validates this starting point, is detailed in the original DiffSinger paper. [4]

It should be noted that the successful implementation of this shallow diffusion hinges on the precise selection of the diffusion step k. While the original paper employs a KL-divergence based technique to estimate the optimal boundary, we approach this estimation with a degree of caution due to potential concerns about the robustness of this method. Recognizing the potential limitations of this method, we undertake a thorough examination through a series of experimental setups. These experiments vary the diffusion steps to critically assess their influence on the DiffSinger model's performance. Our goal is to optimize the use of diffusion mechanisms, ensuring they contribute positively to the synthesis quality and computational efficiency in our singing voice synthesis tasks.

2.4. Dynamic Diffusion Boundary Detection

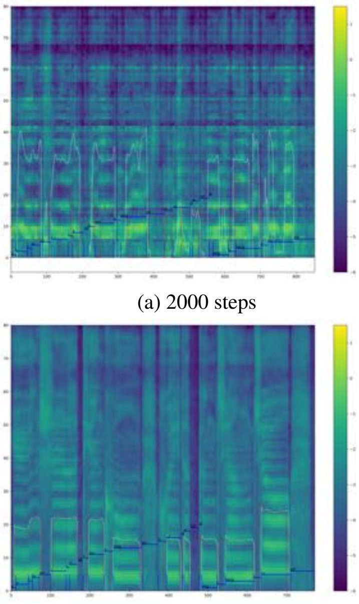
The shallow diffusion mechanism successfully mitigates the trouble of overfitting, as we will show in the experiment section. However, we are not convinced that a constant, hyperparameterized diffusion step k. The value of an optimized k depends on not only the size and overall quality of the training dataset and the other hyperparameters, but also the features of individual front-end output $\tilde{M}$, the current training step, etc. Therefore, we will devise a dynamic boundary detector to compute k at each training step. We will test the performance of our detector against DiffSinger models using constant k configurations on Opencpop dataset, based on a newly designed evaluation suite.

3. Experiments

3.1. Model Adoption

The original DiffSinger code repository has not been maintained and updated, which led us to use a forked version developed by OpenVPI team. The OpenVPI DiffSinger is compatible with Opencpop at the training stage, in which the model trains on batches of audio segments annotated by an integrated transcription file containing phoneme sequence and duration data. The model relies on pitch prediction from Parselmouth. The OpenVPI DiffSinger consists of two individual models. The first model is an acoustic model which learns the acoustic features of the singer(s) and outputs a mel-spectrogram to be converted to a waveform through HiFi-GAN [3], a pretrained vocoder. The second model is a variance model which learns to generate additional parameters for the acoustic model based on customized user input. In this study, we will only train and evaluate the acoustic model, which will be referred to as "the model" by default, for it is the core of singing voice generation.

We trained three models with shallow diffusion steps k at 54, 150, and 400. $k = 54$ is the choice of the original DiffSinger which had been trained on PopCS [4] dataset. 150 and 400 are at the lower and upper bound of the recommended k range respectively. After this, we will train the model with the dynamic boundary detector, and will report the results in the final stage. We will analyze the models according to our evaluation metrics, as well as training and validation loss.


(a) 2000 steps

(b) 160000 steps

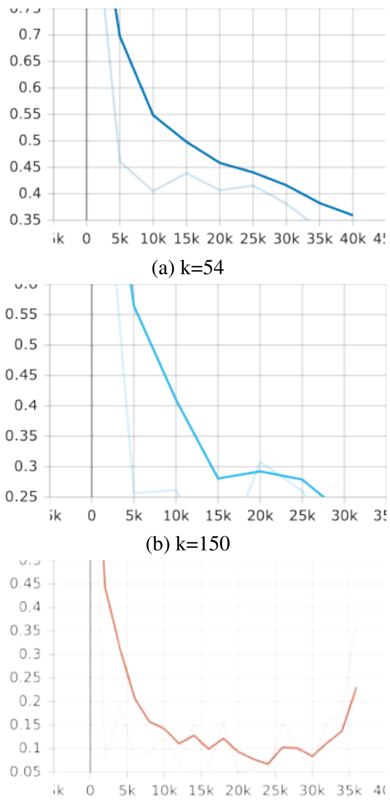
Figure 1. Inference results with different steps

For the purpose of inference, we have reserved 5% of Opencpop dataset. At the inference stage, we feed a DS file ("DS" for "DiffSinger") that contains essential inputs for the model including utterance offset, text, phoneme sequence/duration/number, note sequence/duration/slur, f0 sequence and f0 timestep. The DS file is generated from the ground-truth audio and transcription data. Consequently, We will be able to compare our inference results with the ground-truth mel-spectrograms and audio clips. This evaluation step provides us with more opportunities to quantize the performance of the models.

3.2. Results

The three graphs above represent the validation total loss for three models of different K steps, which means the three different levels of shallow diffusion.

(a) k=54: The graph starts with a high loss value that drops sharply within the first few thousand iterations. After this steep decrease, the loss continues to decrease at a much slower rate. This is typical of a learning process where initial gains are large, and improvements become incremental as the model starts to converge.



(a) k=54

(b) k=150

(c) k=400

Figure 2. Validation Total Loss with different K step

(b) k=150: This graph also starts with a high loss, which decreases sharply but then increases slightly before decreasing again. The second decrease is less steep than the first. The slight increase might indicate a momentary learning challenge or noise in the validation set. However, the overall trend is still downward, suggesting that the model with k=150 is learning over time, albeit with some instability.

(c) k=400: The third graph shows a very different pattern, with the loss starting lower than in the previous two graphs but then increasing significantly, indicating that the model might be diverging or overfitting as the iterations progress. The loss decreases again but remains volatile, with significant ups and downs. This suggests that k=400 is too high and leads to an unstable learning process.

In summary, a lower k value seems to produce a smoother and more stable decrease in validation loss, indicating better learning. As k increases, the model's learning becomes more volatile. The best k value among the three seems to be 54, as it results in the smoothest decrease in loss

3.3. Evaluation

To rigorously assess the performance of our singing voice synthesis model, we introduce a novel evaluation framework that transcends the traditional reliance on subjective metrics: the Mean Opinion Score (MOS) evaluation, which is widely used by previous SVS models, including Diffsinger [4], XiaoiceSing [5], and so on. MOS relies heavily on listeners' subjective perceptions and may lack replicability. Given the variability inherent in human judgment, our objective is to provide a more standardized and replicable suite of evaluation metrics. The proposed framework comprises the following components:

- **Pulse**: Pulse duration is a critical measure reflecting the temporal dynamics of vocal fold vibration. It is integral to the accurate recreation of rhythm and timing in both speech and singing, which are vital for the naturalness of the synthesized voice. By precisely measuring the duration of glottal pulses, we can assess the synthesized voice's ability to replicate the vocal quality of the target, including aspects such as breathiness and tenseness.

- **Formant**: Formants are resonant frequencies of the vocal tract that shape the unique quality of vowels. They are pivotal for speech sound differentiation and can be quantitatively measured. The first two formants, F1 and F2, are typically indicative of vowel sounds. We employ an L1 loss metric to evaluate the accuracy of these formants in the synthesized singing voice, as they closely correlate with perceived vowel quality. We simulate this through artificial vowels created by a click train and bandpass filters, allowing us to objectively measure the model's performance in replicating vocal tract resonances.

- **Pitch**: Pitch accuracy is paramount in singing voice synthesis, reflecting the melodic accuracy of the synthesized voice. We evaluate pitch using both fine-grained deviation measures and overall pitch contour matching to ensure the synthesized voice aligns with the intended melody.

- **Intensity(Dynamics)**: The dynamic range of a singing voice, represented by intensity variations, contributes to the expressiveness of a performance. We measure the intensity of the synthesized voice and compare it to the target to gauge dynamic range accuracy

- **Noise Components**: Noise components, such as sibilance denoted by 'ess' sounds, are inherent in natural speech and singing. They must be accurately modeled to avoid artifacts that detract from the naturalness of the synthesized voice. We assess the presence and quality of these noise components within the synthesized output, ensuring they do not exceed natural levels.

Our evaluation framework, thus, provides a comprehensive and objective method for assessing singing voice synthesis systems. It is designed to give a nuanced picture of performance across temporal, spectral, and dynamic aspects, moving beyond the subjectivity of MOS towards a more definitive and quantitative analysis.

4. Conclusion

In our exploration of the DiffSinger model within the domain of singing voice synthesis, we have introduced a novel evaluation framework, aiming to surpass the limitations of subjective assessment methods like the Mean Opinion Score (MOS). By integrating quantitative metrics into our framework, we evaluated the synthesis quality of our model variants across different diffusion steps, yielding insights into the trade-offs between model complexity and performance.

Our experiments confirmed the efficacy of the shallow diffusion mechanism in capturing the nuances of human singing. Furthermore, our results have paved the way for future research to refine the boundary prediction mechanism and optimize the number of diffusion steps, contributing to the advancement of singing voice synthesis technology.

References

[1] Yin-Ping Cho, Fu-Rong Yang, Yung-Chuan Chang, Ching-Ting Cheng, Xiao-Han Wang, and Yi-Wen Liu. A survey on recent deep learning-driven singing voice synthesis systems, 2021. 1

[2] Jonathon Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. 3

[4] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism, 2022. 1, 2, 3, 4

[5] Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou. Xiaoicesing: A high-quality and integrated singing voice synthesis system, 2020. 4

[6] Jie Wu and Jian Luan. Adversarially trained multi-singer sequence-to-sequence singing synthesizer, 2020. 1