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

Abstract

The field of Singing Voice Synthesis (SVS) features recent models leveraging advanced diffusion techniques to generate realistic singing voices. Our study builds upon the unique shallow diffusion mechanism in DiffSinger, a mainstream open-source SVS model, and offers a computational-linguistic view to quantize and evaluate model generations, in an attempt to optimize synthesis quality and computational efficiency. We have trained models with different configurations of shallow diffusion boundaries, including fixed bounds (the traditional implementation), standard diffusion (a.k.a. full), and newly proposed Adaptive K Schedules. Our evaluation metrics revealed that the standard diffusion model is far from the best although it has the lowest validation loss at the surface, and the adaptive models outperform most fix-bound ones. The new evaluation framework that incorporates quantitative metrics overcome the subjectivity of traditional evaluation methods such as Mean Opinion Score (MOS). Our results have paved the way for future research to refine the boundary prediction mechanism in shallow diffusion.

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 [9]. The task requires an acoustic model to generate key features, which are embedded by visual representations such as a mel-spectrogram [2], 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 [7].

DiffSinger [7] 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 experimented with various implementation decisions in the DiffSinger model and designed an evaluation framework to replace the Mean Opinion Score used in the original study. We committed a few modifications to the boundary prediction mechanism to dynamically handle the intersection of diffusion and reverse trajectories. We then designed an evaluation scheme to bridge the gap between objective measurements and subjective qualities that contribute to a convincing singing voice.

We aim to deliver two key outcomes: firstly, an in-depth comparative analysis between differently configured shallow and standard diffusion models to discern their efficiency and performance in singing voice synthesis; secondly, 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. Related Work

The pipeline of SVS usually consists of an acoustic model to generate the acoustic features conditioned on a music scores, and a vocoder to convert the acoustic features to waveform [7]. Before diffusion, GAN was a popular choice for SVS [6] [1]. However, it has been recognized that GAN is unstable and lacks means of regularization [7].

After the debut of diffusion models, their stability and ease of optimization became preferred by computer vision researchers, especially those working on image syn-

thesis [3]. Diffusion models use Markov chains with fixed parameters to deduce implicit probabilities step by step thus generating images from Gaussian noise. DiffSinger takes advantage of diffusion to generate mel-spectrogram images of high fidelity which can be converted to vivid audio segments.

3. Methodology

3.1. 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.

3.2. Diffusion

We investigate the integration of the diffusion probabilistic model into the SVS system, termed Diffsinger (Liu et al. 2022) [7]. Following the framework proposed by Ho, Jain, and Abbeel (2020) [3], 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 [7]; Ho, Jain, and Abbeel 2020 [3]). 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 [7] 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:

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

3.3. Shallow Diffusion

In our study, we adopted the shallow diffusion mechanism proposed in the DiffSinger framework (Liu et al. 2022) [7]. 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. [7]

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.

3.4. Adaptive K Scheduler for Boundary Detection

Through our experiment with different fix-bound k configurations, we noticed that a configuration of low k impedes convergence, while high k is concerned with overfitting. Therefore, we were motivated to develop a new k schedule that starts at a high value (i.e. $K_{shallow}$) and gravitates toward a threshold (i.e. K_{infer}). At the inference stage, the user should use the threshold value as k because diffusion steps to avoid exceeding training depth.

The Adaptive K Scheduler outputs a k value based the loss of the front-end decoded output before the diffusion block in comparison with the ground-truth mel-spectrogram. Intuitively, a high loss indicates that the model requires more steps to reach a noise equilibrium that is equivalent as adding the same steps of noise to the ground-truth mel-spectrogram. The loss is passed into a `tanh` function to clip it down below 1. The hyperparameter α controls the coefficient of the transformed loss. A momentum that describes the trend of loss over training steps is also added in order to facilitate fast k decline at early stages. β controls the updating rate of the momentum and γ sets the initial momentum. The outcome $\tilde{k}_t$ is finally clamped between K_{infer} and $K_{shallow}$.

Algorithm 1 Adaptive K Scheduler

Require: $\alpha > 0, 0 < \beta < 1, \gamma < 0, K_{infer} < K_{shallow}$
  $L_t \leftarrow \mathcal{L}(\theta_t)$
  if $t = 0$ **then**
    $\mu_t \leftarrow \gamma$
    $k_t \leftarrow K_{shallow}$
  else
    $\tau_t \leftarrow \frac{L_t - L_{t-1}}{L_{t-1}}$
    $\mu_t \leftarrow \beta\mu_{t-1} + (1-\beta)\tau$
    $\tilde{k}_t \leftarrow (\alpha \tanh(L_t) + \mu_t)K_{infer}$
    $k_t \leftarrow \text{round}(\text{clamp}(\tilde{k}_t, K_{infer}, K_{shallow}))$
  end if
  return k_t

As an example, ada20 with the configuration of $\alpha = 20$, $\beta = 0.95$, $\gamma = -0.5$ has the a k schedule graphed below:

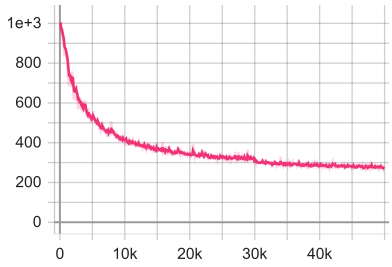
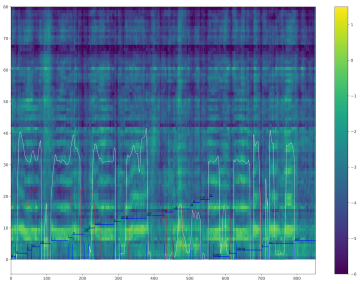


Figure 1. Dynamic k schedule for ada20 model

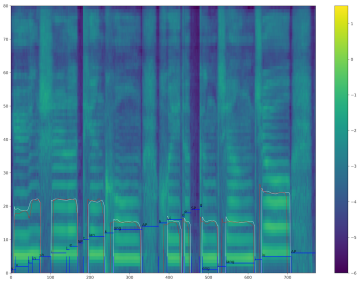
4. Experiments

4.1. Model Adoption

The original DiffSinger code repository has not been maintained and updated, which led us to use a forked version developed by the OpenVPI team. To test DiffSinger, we also used pretrained DiffSinger model to evaluate the inference outcome comparing from different steps in Figure 2. 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 [4]. 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 [5], a pre-trained 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.



(a) After 2000 steps



(b) After 160000 steps

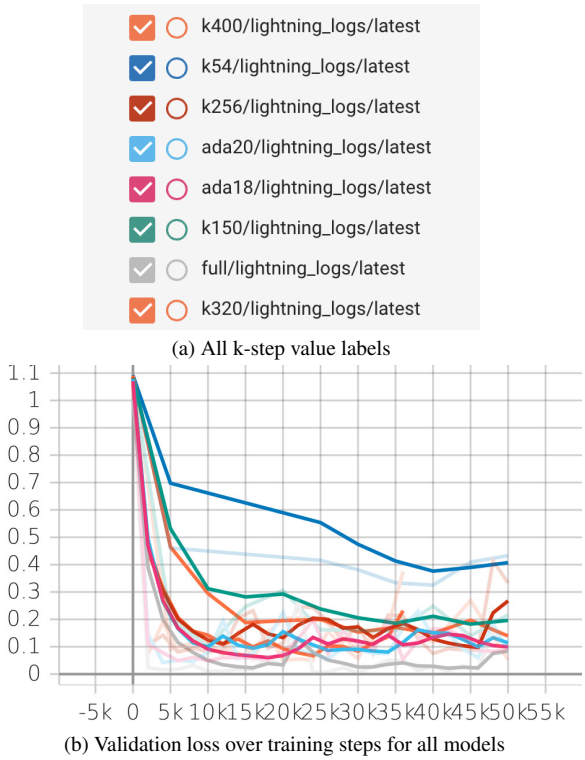
Figure 2. Inference from different steps by the pretrained DiffSinger model

We trained five models with fixed shallow diffusion steps k at 54, 150, 256, 320, 400. $k = 54$ is the choice of the original DiffSinger which had been trained on PopCS [7] 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 our Adaptive K Scheduler at

$\alpha = 18, 20$. We also performed a training $full$ on standard diffusion procedure.

For the purpose of inference, we have reserved 5% of the 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.

4.2. Results



(a) All k-step value labels

(b) Validation loss over training steps for all models

Figure 3. Validation Total Loss with different K-steps

Figure 3 illustrates the total validation loss of models with different K-step boundaries, encompassing traditional fixed bounds($k54, k105, k256, k320, k400$), standard diffusion($full$), and our proposedd Adaptive K Schedulers($ada18, ada20$). Notably, while the standard diffusion model appears to offer the lowest validation loss initially, our findings suggest that this does not balance the performance and efficiency in Singing Voice Synthesis. In fact, models with Adaptive K Scheduler demonstrated a more desirable balance, outperforming most models with fixed K-step boundaries.

As the training steps increased beyond 15k, and models with Adaptive K Scheduler maintained a lower and more stable loss, indicating a robustness in performance that the standard full diffusion model did not achieve. Fixed-bound models, notably $k54$ and $k256$, showed greater variance in their loss, which can indicate the over-fitting.

Furthermore, the sharp initial decrease in validation loss across all models suggests a rapid learning phase, which becomes flat eventually. This flatted effect is notably less happened in adaptive models, pointing towards a more consistent learning and generalization capability.

In conclusion, the Adaptive K Schedule models not only improved upon the traditional fixed-bound methods but also provided a clear path for future advancements in shallow diffusion for SVS.

4.3. 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 [7], XiaoiceSing [8], 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:

- **Mel-spectrogram**: The mel-spectrogram loss serves as a primary indicator of acoustic fidelity. It is computed as L1 Loss between the pixel-wise values of the normalized mel-spectrogram of the generated audio and the ground truth. This metric emphasizes the importance of maintaining the overall acoustic structure in singing voice synthesis.
- **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. To specifically target the clarity of articulation in synthesized vocals, we compute both F1 and F2 formant losses. Utilizing Textgrid annotations, vowels are isolated from the audio data, and an L2 loss is computed between the model's formants and those of the target audio. The aggregate of these losses provides a precise measure of the model's ability to replicate the distinct resonances that characterize vowel sounds.
- **Intensity (Dynamics)**: The expressiveness and power of a singing voice are encapsulated within its dynamic range. We evaluate this aspect through an L2 loss comparison of the normalized intensity contours of each

synthesized audio segment against the ground truth. This metric underscores the model's capacity to produce the dynamic variations that contribute to a more lifelike singing experience.

Our evaluation framework is designed to provide a comprehensive and objective method for assessing singing voice synthesis systems, moving beyond the subjectivity of MOS towards a more definitive and quantitative analysis. The inference performance is assessed based on a linear combination of the aforementioned metrics. This composite score provides a balanced measure that accounts for acoustic quality, clarity of articulation, expressiveness, and computational efficiency, thus enabling a nuanced evaluation of the model variants:

$$\begin{aligned}\mathcal{L}(\hat{Y}, Y) = w_1 \times \mathcal{L}_{mel}(\hat{Y}, Y) \\ + w_2 \times \mathcal{L}_{F1}(\hat{Y}, Y) \\ + w_3 \times \mathcal{L}_{F2}(\hat{Y}, Y) \\ + w_4 \times \mathcal{L}_{Intensity}(\hat{Y}, Y))\end{aligned} \tag{5}$$

where $\mathcal{L}(\hat{Y}, Y)$ denotes the total loss weighted by different metrics, $w_1 = 0.6$, $w_2 = 0.1$, $w_3 = 0.1$, and $w_4 = 0.2$.

The determination of weights for each metric in our evaluation framework was guided by the specific characteristics and relative importance of each metric in capturing the quality of synthesized singing voices. Given the comprehensive coverage of acoustic information by the mel-spectrogram, we assigned it the highest weight of 0.6, reflecting its critical role in accurately representing the overall acoustic structure of the voice segment. In contrast, the formant metrics, which include both F1 and F2 losses, focus specifically on the clarity of vowel articulations. Considering that vowels constitute only a portion of a voice segment and that these metrics collectively measure the quality of these portions, we assigned each a weight of 0.1, summing to 0.2 for the entire formant category. This allocation acknowledges the significance of vowel clarity while recognizing that it is only one component of voice synthesis quality. The intensity metric, assessing the dynamic range and thus the expressiveness of the voice, was given a weight of 0.2. This reflects its importance in distinguishing the dynamic differences between the synthesized and the ground truth voices but acknowledges that the mel-spectrogram provides a broader measure of voice quality.

The convergence of these metrics presents an innovative means to objectively quantify the nuances of synthesized singing, offering a robust alternative to subjective assessment methods. Our evaluation framework not only reveals the subtleties captured by the model but also guides future optimization of the diffusion process.

The results of our evaluation provide compelling evidence of the effectiveness of our proposed metrics in objectively assessing the quality of synthesized singing voices:

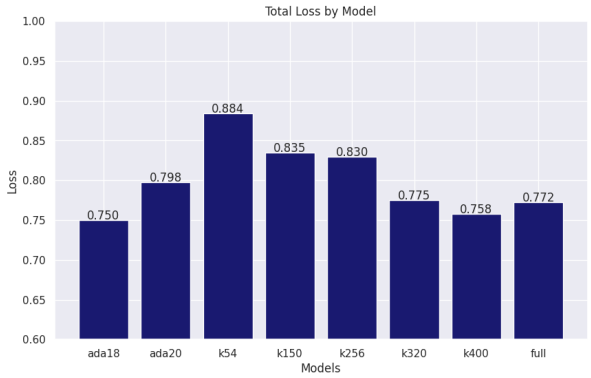


Figure 4. Total loss evaluation by our evaluation metrics

5. Conclusion

This paper presented two principal contributions that address critical challenges in the field of singing voice synthesis. First, we introduced the Adaptive K Scheduler, a novel dynamic boundary detector algorithm that determines the optimal number of diffusion steps (k) during each training iteration. This mechanism is pivotal in balancing synthesis quality against computational efficiency. Our empirical results demonstrate that this approach not only maintains low loss values indicative of high synthesis quality but also enhances computational efficiency.

Our second major contribution is the introduction of novel evaluation metrics that offer quantitative methods to assess the quality of voice synthesis tasks. By moving away from traditional subjective assessments like the Mean Opinion Score (MOS), these metrics provide a robust quantitative methodology for evaluating the quality of synthesized voices. These metrics have been meticulously designed to capture a comprehensive range of acoustic properties, ensuring a thorough and objective assessment.

Together, these contributions signify a substantial advancement in the development and evaluation of voice synthesis systems. The Adaptive K Scheduler opens a new pathway of efficient model training, which could be integrated into the current DiffSinger architecture as an viable option with a flexible hyperparameter lineup. Our evaluation metrics establish a new standard for objective SVS quality assessment, for which there is no precedent studies as alternatives. It has the potential to become a universal benchmark for SVS experiments.

Looking ahead, our future work will expand on these foundations in several ways. We plan to conduct user studies to ensure that our quantitative metrics correlate well with subjective quality assessments, as measured by Mean Opinion Scores. We hypothesize that voices rated higher by our evaluation framework will also receive higher subjective scores, reinforcing the validity of our approach.

References

[1] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu. Hifisinger: Towards high-fidelity neural singing voice synthesis, 2020. 1

[2] 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

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

[4] Yannick Jadoul, Bill Thompson, and Bart de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. 3

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

[6] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee. Adversarially trained end-to-end korean singing voice synthesis system, 2019. 1

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

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

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