INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Prior-free Guided TTS: An Improved and Efficient Diffusion-based Text-Guided Speech Synthesis

Won-Gook Choi, So-Jeong Kim, TaeHo Kim, and Joon-Hyuk Chang

Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
onlyworld94@hanyang.ac.kr, sj2021162470@hanyang.ac.kr, kth.address@gmail.com, jchang@hanyang.ac.kr

Abstract

Recently, diffusion models have exhibited higher sample quality with guidance, such as classifier guidance and classifier-free guidance. However, these guidances have limitations: they require extra classifiers or joint training, and incur additional sampling cost. In this study, we propose prior-free guidance diffusion model and prior-free guided text-to-speech (PfGuided-TTS) that can generate a speech at a quality as high as other guidances without extra training resources and computational cost. PfGuided-TTS can generate higher human perceptual quality speech than the existing autoregressive (AR) and non-AR models, including diffusion-based TTS on LJSpeech. In addition, we provide a schematic describing why and how classifier- and prior-free guided scores produce high-fidelity samples.

Index Terms: text-to-speech, diffusion model, guided score

1. Introduction

Most neural speech syntheses comprise a text-to-speech (TTS) model and a vocoder. TTS encodes text into text-embedding vectors and decodes the embeddings into acoustic features such as a log-mel spectrogram. Subsequently, the vocoder converts the acoustic features into an audible audio signal. This study explores an acoustic model composing a neural TTS.

The two main approaches for designing neural TTS are autoregressive (AR) and non-AR models. The AR models, such as Tacotron2 [1], generate high-quality samples by synthesizing frame-by-frame sequentially. Compared to AR models, the non-AR models, such as FastSpeech1, 2 [2, 3] and Glow-TTS [4], have exhibited as good performance as the AR models with lesser computational cost (especially for long sequences). Recently, a denoising diffusion probabilistic model (DDPM) [5], a non-AR model, has also successfully achieved high quality speech and pitch variety in TTS [6–8]. For instance, Jeong *et al.* [6] proposed Diff-TTS that outperformed other AR and non-AR models with fewer parameters. They also demonstrated that users can control the pitch diversity by scaling the noise temperature and the trade-off between the sample quality and inference speed by adjusting the sampling steps.

As diffusion models have been developed, truncation-like methods for diffusion models, termed *guidance* [9, 10] that can improve the sample quality (while sacrificing diversity), have been investigated. Dhariwal *et al.* [9] introduced *classifier guidance* that can control the trade-off between sample quality and diversity. Instead of denoising the estimated scores of the conditional diffusion model, they adopted the unconditional diffusion model's scores guided by the gradients of the pre-trained classifier. Ho *et al.* [10] derived *classifier-free guidance* from classifier guidance, in which the guided scores are calculated as a combination of the conditional and unconditional models' scores. To avoid extra training, they joint-trained the conditional and unconditional models by defining the null class.

Although these guidances successfully enhanced the sample quality, limitations remain. First, the classifier-guided score requires the separate classifier and it could demand high training resources. For example, Kim *et al.* [8] trained a phoneme classifier for Guided-TTS using 982 h extra utterances. Also, computation for the gradients of the classifier or scores of the unconditional model incurs additional sampling cost. This limitation is critical since DDPM itself requires high computational cost for sampling.

In this study, we propose *prior-free guidance* and *prior-free guided TTS* (PfGuided-TTS) that can synthesize higher quality speech while requiring neither a pre-trained model nor extra computation. Instead of mixing the estimated scores of the unconditional model, prior-free guidance uses the score of the probability distribution of the forward process. The forward process does not require any neural network; hence, the sampling cost is sustained and joint training is avoided. Experimental results show that PfGuided-TTS can generate higher human perceptual quality samples than Diff-TTS and even classifier-free guided TTS (CfGuided-TTS), with a similar computational cost as the normal sampling.

This paper contains how prior-free guidance is induced from the classifier-free guidance and a toy example in Section 3, and experimental results in Section 5. Also, in Section 6, we will discuss why and how the guided scores can enhance the sample quality.

2. Backgrounds

2.1. Diffusion model-based TTS

DDPM [5] is a method for designing a generative model, which consists of a pre-defined diffusion process and a parameterized denoising process. Let X_0 and y denote the log-mel spectrogram of a speech and its corresponding text, respectively. A TTS model $p_\theta(X_0|y)$ can be factorized as follows:

$$p_\theta(X_0|y) = \int p_\theta(X_{0:T}|y)dX_{1:T}, \tag{1}$$

where T is the total diffusion timesteps and $\{X_t\}_{t=1}^{T}$ are latent variables of X_0. DDPM aims to model the T steps *denoising process*, $p_\theta(X_{0:T}|y)$, and the process is assumed to be a Markov chain starting from $X_T \sim \mathcal{N}(0, I)$: $p_\theta(X_{0:T}|y) := p(X_T)\prod_{t=1}^{T} p_\theta(X_{t-1}|X_t, y)$ where $p_\theta(X_{t-1}|X_t, y) := \mathcal{N}(X_{t-1}; \mu_\theta, \Sigma_\theta)$. On the other hand, the *diffusion process* is defined as a pre-fixed Gaussian distribution with a Markov chain $q(X_{1:T}|X_0) := \prod_{t=1}^{T} q(X_t|X_{t-1})$, where $q(X_t|X_{t-1}) := \mathcal{N}(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I)$. $\{\beta_t\}_{t=1}^{T}$ is a variance schedule that corrupts data X_0 more strongly as

10.21437/Interspeech.2023-506

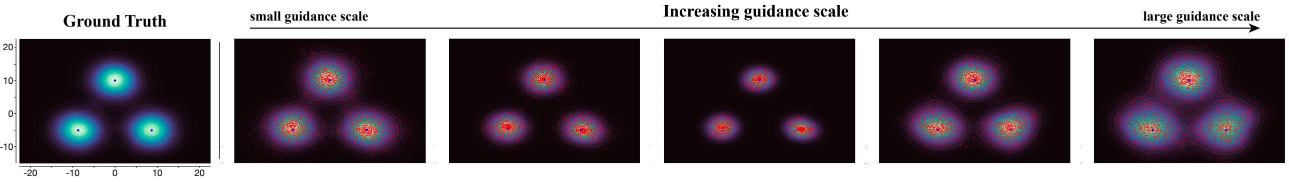


Figure 1: *Toy 2D examples for the prior-free guidance on the three Gaussian mixture distribution.*

the step progresses. With the notations $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$, X_t can be sampled from X_0: $q(X_t|X_0) := \mathcal{N}(X_t; \sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I)$. The neural network is optimized by minimizing the following simplified objective:

$$L_t = \mathbb{E}_{X_0,\epsilon}[||\epsilon_t - \epsilon_\theta(X_t, y, t)||^2], \tag{2}$$

where the output of the network ϵ_θ estimates the Gaussian noise ϵ_t on X_t (also known as the score estimation [11, 12]). Finally, X_{t-1} is sampled from $p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_{\theta,t}, \sigma_t^2 I)$ where $\mu_{\theta,t} = \frac{1}{\sqrt{\alpha_t}}(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta)$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. If the process applies the noise temperature η, $X_{t-1} = \mu_{\theta,t}X_t + \eta\sigma_t z_t$, where $z_t \sim \mathcal{N}(0, I)$. When $\eta < 1$, blurry images are generated in the image synthesis [9], and the pitch variability decreases in speech synthesis [6]. We will discuss how the proposed guidance and noise temperature affect the pitch diversity in Section 5; but basically, we set $\eta = 1$ for all experiments.

2.2. Guided diffusion models

Classifier guidance [9] tunes the denoising score to perform a truncation-like effect on sampling. Dhariwal *et al.* observed that a conditional diffusion model can be factorized into an unconditional diffusion model and classifier as $p_\theta(X_{t-1}|X_t, y) \propto p_\theta(X_{t-1}|X_t)p_\theta(y|X_{t-1})$. The denoising score on step t is modified by scaling the distribution of the classifier as follows:

$$\tilde{\epsilon}_\gamma(X_t, y, t) = \epsilon_\theta(X_t, t) + \gamma\sigma_t\nabla_{X_t}\text{log}p_\theta(y|X_t), \tag{3}$$

where γ denotes the guidance scaling. If we have an extra classifier targeting the conditioning information, we can sample high-quality but low-diversity data when $\gamma > 1$.

Ho *et al.* [10] proposed **classifier-free guidance** from classifier guidance by factorizing the classifier $p_\theta(y|X_{t-1})$ into $p_\theta(X_{t-1}|X_t, y)p_\theta(y|X_t)/p_\theta(X_{t-1}|X_t)$.

$$\begin{aligned} \tilde{p}_\gamma(X_{t-1}|X_t, y) &\propto p_\theta(X_{t-1}|X_t)p_\theta(y|X_{t-1})^\gamma \\ &\propto p_\theta(X_{t-1}|X_t)^{1-\gamma}p_\theta(X_{t-1}|X_t, y)^\gamma. \end{aligned} \tag{4}$$

Using this property, the guided denoising score is expressed as:

$$\tilde{\epsilon}_\gamma(X_t, y, t) = (1-\gamma)\epsilon_\theta(X_t, y = \phi, t) + \gamma\epsilon_\theta(X_t, y, t). \tag{5}$$

Here, ϕ denotes the null condition, which is a technique to avoid training the models separately (i.e., a technique for joint training). This guidance does not require any classifier; thus, both training and sampling processes are simplified.

3. Proposed method

Although classifier-free guidance was successfully achieved in a simple form with similar effects to classifier guidance, it still has limitations [10]. The sampling process costs twice as much as the normal sampling because both conditioned and unconditioned scores must be computed. In this section, we introduce the costless-guided denoising score, named *prior-free guidance*.

Let us recall $p_\theta(X_{t-1}|X_t)$ in Eq. (4), which can be factorized into $p_\theta(X_t|X_{t-1})p_\theta(X_{t-1})/p_\theta(X_t)$ using Bayes' rule. If T is sufficiently large, the diffusion rates $\{\beta_t\}_{t=1}^T$ are relatively close to zero [13] to prevent rapid diffusion. Then, we can assume that $p_\theta(X_t) \approx p_\theta(X_{t-1})$, which indicates that the means and variances of the diffused data adjacent step are almost the same (this assumption is more convincing when t is closer to T). Subsequently, since $p_\theta(X_t|X_{t-1})$ is the forward process, it becomes the normal distribution $\mathcal{N}(\sqrt{1-\beta_t}X_{t-1}, \beta_t I)$ regardless of parameters θ. Thus, Eq. (4) can be expressed as:

$$\tilde{p}_\gamma(X_{t-1}|X_t, y) \propto p(X_t|X_{t-1})^{1-\gamma}p_\theta(X_{t-1}|X_t, y)^\gamma. \tag{6}$$

Then, the score function for $\tilde{p}_\gamma(X_{t-1}|X_t, y)$ is derived as:

$$\begin{aligned} &\nabla_{X_{t-1}}\text{log}\tilde{p}_\gamma(X_{t-1}|X_t, y) \\ &= \nabla_{X_{t-1}}[(1-\gamma)\text{log}p(X_t|X_{t-1}) + \gamma\text{log}p_\theta(X_{t-1}|X_t, y)] \\ &= (1-\gamma)\frac{1}{\sqrt{\beta_t}}\epsilon_t - \gamma\frac{1}{\sigma_t}\epsilon_\theta(X_t, y, t). \end{aligned} \tag{7}$$

Note that $\epsilon_t \equiv -\epsilon_t$; therefore, the modified-guided denoising score is defined as:

$$\tilde{\epsilon}_\gamma(X_t, y, t) := (1-\gamma)\frac{\sigma_t}{\sqrt{\beta_t}}\epsilon_t + \gamma\epsilon_\theta(X_t, y, t), \tag{8}$$

then, $\mu_{\theta,t}$ is modified into $\tilde{\mu}_{\gamma,t} = \frac{1}{\sqrt{\alpha_t}}(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_\gamma(X_t, y))$.

Figure 1 shows a toy 2D example describing the effect of numerically induced prior-free guidance on various conditioning scales. As the conditioning scales increase enough, the samples exhibit greater convergence to the mean of each distribution. However, if the scale becomes too high, the samples are farther away from each mean. This pattern occurs equally in speech synthesis: if γ is large enough, the synthesized speech quality improved, but if it is too large, the quality deteriorates. This phenomenon will be explained in detail in Section 6.

One curious fact is that the classifier- and prior-free guidances are defined under Bayes' rule by replacing the conditional or unconditional model with an unconditional model or a normal distribution, respectively. However, the approach of reversing a probability distribution using Bayes' rule is not always valid, even if it results in a similar effect that can trade off sample quality and diversity (this fact is also mentioned in [10]). Instead, we will provide a schematic analysis of the expected trajectory of the sampling process in Section 6, explaining why both classifier- and prior-free guidances generate better samples than the normal sampling process.

In the following sections, we will compare prior-free guidance to classifier-free guidance but classifier guidance since classifier guidance requires too high training resources. We consider that it is sufficient to verify that our method has an efficient sampling process while generating high-quality speech in comparison with classifier-free guidance alone.

4. Experiments

4.1. Dataset and experimental setup

We trained TTS models using LJSpeech [14] recorded by a single female speaker at a sampling rate of 22.05 kHz. It consists of 13,100 audio clips whose length varied from 1-10 s with a total duration of 24 h. We randomly split the data into training, validation, and test set into 12,500, 100, and 500 clips, respectively. The log-mel spectrogram was extracted with 1024 FFT and window sizes, 256 hop sizes, and 80 mel bins. The audible signal was synthesized using the pre-trained vocoder, HiFi-GAN [15]. We optimized the network for 1 M iterations using the Adam optimizer [16], whose learning rates and weight decays were both 0.0001 on an NVIDIA 3090 GPU. We provide our demo page containing synthesized samples.[1]

4.2. Neural networks and diffusion process

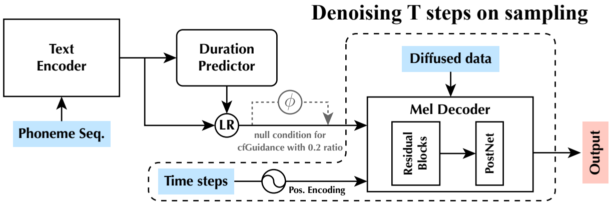


Figure 2: *Overall network architecture for the experiments (Diff-TTS, CfGuided-TTS and PfGuided-TTS in Table 1).*

The network architecture and configurations were the same as those of Diff-TTS in [6], except for the PostNet, which composes the mel decoder (Fig. 2). We investigated that the CBHG (1D-Conv. Banks + Highway + bi-GRU) used in Tacotron's PostNet [17] worked well for the mel decoding than the PostNet used in Tacotron2 [1]; thus we used the same architecture of CBHG for the PostNet. For the details, the input dimension, size of K, and projections for the convolutional and highway layer were set to 512, 8, 256, and 128, respectively. Subsequently, the output dimension was linearly projected on the number of mel bins. We used the diffusion steps $T = 200$ and the linearly spaced variance schedule $\beta_t \in [5 \times 10^{-4}, 0.1]$. For training the CfGuided-TTS, we replaced all the elements of the text features with 0.01 for null conditioning with a 20% ratio. For training the duration predictor, we used the Montreal forced aligner [3]. Finally, the total number of parameters was 13.4 M.

4.3. Accelerated sampling

We adopted the accelerated sampling algorithm of [6] that was motivated by [18]. Let S and $X_{t'_1:t'_S}$ denote the number of accelerated sampling steps and the subsequence of the latent variables $X_{1:T}$, respectively (Fig. 3). We set $S = \lceil T/\tau + 1 \rceil$, where $\lceil \cdot \rceil$ and τ denote the ceiling function and decimation factor, respectively. We set $t'_s = T - (S - s)\tau$ for integers $1 < s \leq S$ and $t'_1 = 1$. Details for the accelerated sampling algorithm, it is modified into:

$$X_{t'_{s-1}} = \sqrt{\frac{\bar{\alpha}_{t'_{s-1}}}{\bar{\alpha}_{t'_s}}}(X_{t'_s} - \sqrt{1 - \bar{\alpha}_{t'_s}}\epsilon_\theta(X_{t'_s}, y, t)) + \sqrt{1 - \bar{\alpha}_{t'_{s-1}} - \sigma^2_{t'_s}}\epsilon_\theta(X_{t'_s}, y, t) + \eta\sigma_{t'_s} z_{t'_s}, \tag{9}$$

where $\sigma_{t'_s} = \sqrt{\frac{1-\bar{\alpha}_{t'_{s-1}}}{1-\bar{\alpha}_{t'_s}}\beta_{t'_s}}$. For $s = 1$, t'_0 is 0.

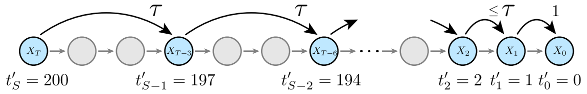


Figure 3: *A graphical example for the accelerated sampling process when $\tau = 3$. Gray circles depict the skipped steps. Note that the step size could be smaller than τ for $s = 2$ when $T - 1$ is not divisible by τ.*

4.4. Evaluation

To evaluate the audio quality of TTS models, we used 5-scale mean opinion score (MOS) tests with a 95% confidence interval (CI) on 15 participants for subjective evaluation, and measured a character error rate (CER) using the pre-trained automatic speech recognition model from the NEMO toolkit[2] for objective evaluation. In addition, we measured real-time factors (RTF) on a single NVIDIA 3090 GPU to evaluate the inference speed. Both the CER and RTF were measured for 500 sentences composing the test set, and the MOS tests were performed for 30 and 20 sentences randomly chosen from the test set for the experiments in Table 1 and 2, respectively.

5. Results

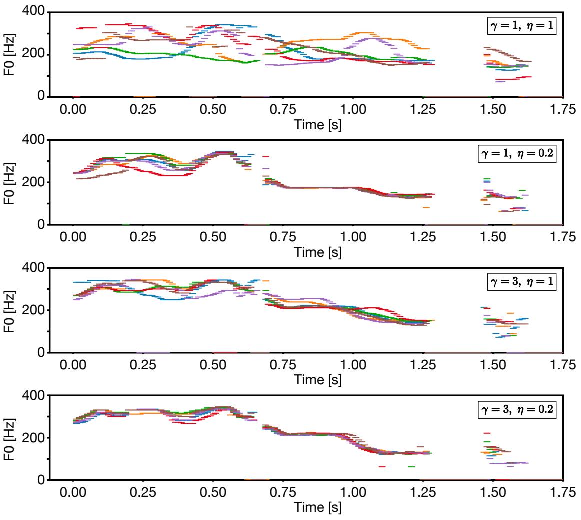


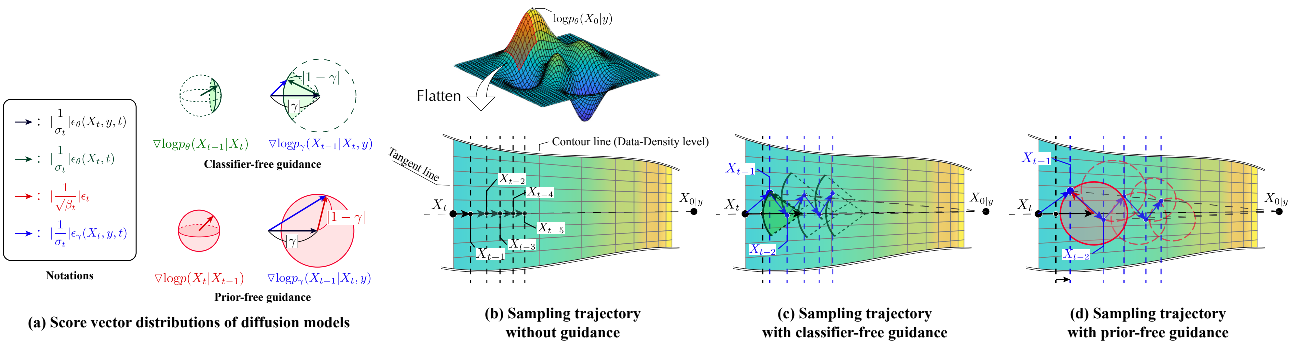
Figure 4: *Pitch tracks on different γ and η for the 6 differently generated speech samples from the same sentence.*

As shown in Table 1, both Cf and PfGuided-TTS generated the higher quality audio than Diff-TTS. Interestingly, the clarity (CER) was highest on CfGuided-TTS, but the human perceptual quality and naturalness (MOS) were highest on PfGuided-TTS with the lower computational cost; even outperformed the speeches vocoded from ground truth mel. In addition, the audio qualities showed robust on the accelerated sampling when $\tau = 4$. Although the audio quality degraded when $\tau = 10$, MOS was still higher than other TTS models and RTF enhanced a lot. Here, the optimal γ decreased with increasing τ for the same reason that will be discussed in Section 6.

To investigate how the guidance scale and noise temperature affect the prosody variety, we extracted the pitch contours with different γ and η. As already explored in [6], lower values of η, correspond to lower diversity. However, PfGuided-TTS generated speeches with more diverse prosodies than controlling the noise temperature. The toy example in Fig. 1 shows

[1] https://wgk.notion.site/Demo-Prior-free-Guided-TTS-76c430174bc14df59fa077ba19b08bf2

[2] https://huggingface.co/nvidia/stt_en_conformer_ctc_large


$\log p_\theta(X_0|y)$
Flatten
$: |\frac{1}{\sigma_t}|\epsilon_\theta(X_t,y,t)$
$: |\frac{1}{\sigma_t}|\epsilon_\theta(X_t,t)$
$: |\frac{1}{\sqrt{\beta_t}}|\epsilon_t$
$: |\frac{1}{\sigma_t}|\epsilon_\gamma(X_t,y,t)$
Notations
$|1-\gamma|$
$|\gamma|$
$\nabla \log p_\theta(X_{t-1}|X_t)$
$\nabla \log p_\gamma(X_{t-1}|X_t,y)$
Classifier-free guidance
$\nabla \log p(X_t|X_{t-1})$
$\nabla \log p_\gamma(X_{t-1}|X_t,y)$
Prior-free guidance
Tangent line
Contour line (Data-Density level)
X_t
X_{t-1}
X_{t-2}
X_{t-3}
X_{t-4}
X_{t-5}
$X_{0|y}$

(a) Score vector distributions of diffusion models

(b) Sampling trajectory without guidance

(c) Sampling trajectory with classifier-free guidance

(d) Sampling trajectory with prior-free guidance

Figure 5: *Schematic diagram of the sampling trajectories with and without guidances. We denote the notation $\nabla_{X_{t-1}}(\cdot) = \nabla(\cdot)$ for brevity. All the score vectors are random variable but for simplification, we assume that $\nabla \log p_\theta(X_{t-1}|X_t, y)$ always points in the direction for the shortest path.* ***(a)*** *illustrates the distributions of score vectors. We express the $\nabla \log p_\theta(X_{t-1}|X_t)$, and $\nabla \log p(X_t|X_{t-1})$ as 3D vectors, but $\nabla \log \tilde{p}_\gamma(X_{t-1}|X_t, y)$ as 2D vectors for simplification.* ***(b-d)*** *illustrate the sampling trajectories on the simplified data distribution. Note that $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t} < \sqrt{\beta_t}$ for all t, thus guided sampling is always ahead of normal sampling in (d).*

Table 1: *Comparison of the different TTS models. Note that the Diff-TTS is same to the $\gamma = 1$ for the PfGuided-TTS. We first chose the optimal $\gamma = 3$ under CER, but after the MOS test shown in Table 2, we found that the optimal γ also could be 2 respect to MOS.*

Method	**MOS (CI)**	**CER(%)**	**RTF**
GT	$4.46 \pm .044$	1.01	–
GT (Mel + HiFiGAN)	$4.23 \pm .053$	1.27	–
Tacotron2	$4.13 \pm .053$	2.78	0.055
Glow-TTS ($T = 0.333$)	$4.07 \pm .056$	3.83	**0.005**
Diff-TTS	$3.94 \pm .058$	3.37	0.337
CfGuided-TTS ($\gamma = 4.5$)	$4.10 \pm .057$	**2.75**	0.656
PfGuided-TTS ($\gamma = 3$)	$\mathbf{4.39 \pm .044}$	2.82	0.337
$\tau = 4, \ \ \gamma = 2.5$	$4.36 \pm .044$	2.83	0.087
$\tau = 10, \gamma = 2$	$4.25 \pm .045$	2.89	0.036
$\tau = 25, \gamma = 1.2$	$4.05 \pm .054$	2.88	0.017

Table 2: *Comparison of the audio quality by the different guidance scales of PfGuided-TTS.*

Scale (γ)	1	2	3	5	7	10
MOS (CI)	3.98 ±.06	**4.49** **±.04**	4.40 ±.04	4.26 ±.05	2.97 ±.06	1.15 ±.02
CER (%)	3.37	2.91	**2.82**	3.00	3.21	4.49

that sample variance increased as γ increased. But, in practice, controlling the diversity by modulating γ as a higher value is difficult since the audio quality deteriorates drastically as γ becomes too higher (Table 2).

6. Discussions

In this section, we provide a schematic analysis to discuss how the classifier- and prior-free guidance affect the sampling trajectory. Figure 5 depicts the log-likelihood distribution of data X_0 and expresses all score vectors of the distributions $\tilde{p}_\gamma(X_{t-1}|X_t, y), p_\theta(X_{t-1}|X_t, y), p_\theta(X_{t-1}|X_t)$, and $p(X_t|X_{t-1})$ as 2D or 3D vectors. Note that $\gamma > 1$ for guidance, therefore $1 - \gamma < 0$; thus, the direction of the vector is reversed (Fig. 5 a).

The direction of $\nabla \log p(X_t|X_{t-1})$ is unbiased (i.e. completely random), but the direction of $\nabla \log p_\theta(X_{t-1}|X_t)$ is biased to the data point $X_{0|y}$ since $p_\theta(X_{t-1}|X_t)$ is trained to maximize $p_\theta(X_0)$. During normal sampling (Fig. 5 b), X_{t-1} reaches a relatively low density level (black dashed vertical line on Fig. 5 b-d). However, if the score is guided, X_{t-1} could reach the higher density level than normal sampling (blue dashed-line on Fig. 5 c, d). With respect to diversity, the smaller the variance of the Gaussian distribution is (i.e., low-diversity distribution), the steeper the gradient of the log-likelihood is (samples reach a higher density level per step). For these reasons, the guided scores generate high-quality but low-diversity samples; and if γ increases, the phenomenon is strengthened.

However, one question remains: why did the toy example in Fig. 1 and the experimental results in Table 2 exhibited lower sample quality when γ was too large? We described the trajectory with a simple data distribution model for the analysis in Fig. 5, but in practice, data has higher dimensions and more complex distribution. In addition, we mentioned that the direction of $\nabla \log p(X_t|X_{t-1})$ was unbiased. Therefore, if γ is set too large, X_{t-1} could lose the way to the mode of $X_{0|y}$ (this is similar to why a very high learning rate is not set when optimizing the neural network). The same reason applies for the accelerated sampling; the magnitude of the vector increases with $\tau > 1$, hence γ should be smaller than that of $\tau = 1$. Therefore, when sampling data using the proposed prior-free guidance, γ should be carefully set depending on the tasks and total denoising steps.

7. Conclusion

In this study, we introduced prior-free guidance, which generates high-quality samples without increasing computational cost by numerically substituting the unconditional reverse process with the forward process. PfGuided-TTS synthesized high-quality speech while preserving the computational cost and keeping high quality despite accelerated sampling. In addition, using a schematic, we explained how the guided scores can generate a better sample quality. The guidance made the sample reach a higher density level, but too large scale, the sample quality deteriorated.

8. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation)

9. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2018, pp. 4779–4783.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 32, 2019.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations* (*ICLR*), 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[4] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/5c3b99e8f92532e5ad1556e53ceea00c-Abstract.html

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 6840–6851.

[6] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A Denoising Diffusion Model for Text-to-Speech," in *Proc. Interspeech*, 2021, pp. 3605–3609.

[7] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proc. International Conference on Machine Learning* (*ICML*). PMLR, 2021, pp. 8599–8608.

[8] H. Kim, S. Kim, and S. Yoon, "Guided-TTS: A diffusion model for text-to-speech via classifier guidance," in *Proc. International Conference on Machine Learning* (*ICML*), vol. 162. PMLR, 2022, pp. 11 119–11 133.

[9] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 34, 2021, pp. 8780–8794.

[10] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[11] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 11 287–11 302.

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. International Conference on Learning Representations* (*ICLR*), 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

[13] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. International Conference on Machine Learning* (*ICML*). PMLR, 2015, pp. 2256–2265.

[14] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[15] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 17 022–17 033.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[18] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. International Conference on Learning Representations* (*ICLR*), 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP