Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance

Heeseung Kim [*1] **Sungwon Kim** [*1] **Sungroh Yoon** [1 2]

Abstract

We propose Guided-TTS, a high-quality text-to-speech (TTS) model that does not require any transcript of target speaker using classifier guidance. Guided-TTS combines an unconditional diffusion probabilistic model with a separately trained phoneme classifier for classifier guidance. Our unconditional diffusion model learns to generate speech without any context from untranscribed speech data. For TTS synthesis, we guide the generative process of the diffusion model with a phoneme classifier trained on a large-scale speech recognition dataset. We present a norm-based scaling method that reduces the pronunciation errors of classifier guidance in Guided-TTS. We show that Guided-TTS achieves a performance comparable to that of the state-of-the-art TTS model, Grad-TTS, without any transcript for LJSpeech. We further demonstrate that Guided-TTS performs well on diverse datasets including a long-form untranscribed dataset.

1. Introduction

Neural text-to-speech (TTS) models have been achieved to generate high-quality human-like speech from given text (van den Oord et al., 2016; Shen et al., 2018). In general, TTS models are conditional generative models that encode text into a hidden representation and generate speech from the encoded representation. Early TTS models are autoregressive generative models that generate high-quality speech but suffer from a slow synthesis speed due to the sequential sampling procedure (Shen et al., 2018; Li et al., 2019). Owing to the development of non-autoregressive generative models, recent TTS models can generate high-quality speech with faster inference speed (Ren et al., 2019; 2021; Kim et al., 2020; Popov et al., 2021). Recently, high-quality end-to-end TTS models have been proposed that generate raw waveforms from text at once (Kim et al., 2021; Weiss et al., 2021; Chen et al., 2021b).

现有的TTS模型需要在转录数据上训练，直接使用未转录数据是个挑战。

Despite the high quality and fast inference speed of speech synthesis, most TTS models can only be trained if the transcribed data of the target speaker are provided. Although long-form untranscribed data, such as audiobooks or podcasts, is available on various websites, it is challenging to use these speech data as training datasets for existing TTS models. To utilize these untranscribed data, long-form untranscribed speech data has to be segmented into sentences, and each segmented speech should then be accurately transcribed. Since the existing TTS models directly model the conditional distribution of speech given text, the direct usage of untranscribed data remains a challenge.

尽管有一些方法使用未转录音频，但这些方法很大程度依赖预训练的multispeaker TTS model

There have also been approaches using untranscribed speech to adapt the pre-trained multi-speaker TTS model for few-shot TTS synthesis (Jia et al., 2018; Yan et al., 2021). These adaptive TTS models rely heavily on a pre-trained multi-speaker TTS model, which is challenging to train and requires high-quality multi-speaker TTS datasets. Also, due to the difficulties of generalization, they underperform in comparison to high-quality single-speaker TTS models such as Glow-TTS and Grad-TTS (Kim et al., 2020; Popov et al., 2021) trained on a large amount of transcribed data.

In this work, we propose Guided-TTS, a high-quality TTS model that learns to generate speech with an unconditional DDPM and performs text-to-speech synthesis using classifier guidance. By introducing a phoneme classifier trained on a large-scale speech recognition dataset, Guided-TTS does not use any transcript of the target speaker for TTS. Trained on untranscribed data, our unconditional diffusion probabilistic model learns to generate mel-spectrograms without context. As the untranscribed data does not have to be aligned with the text sequence, we simply use random chunks of untranscribed speech to train our unconditional generative model. This allows us to build training datasets without extra effort in modeling the speech of speakers for which only long-form untranscribed data is available.

To guide the unconditional DDPM for TTS, we train a frame-wise phoneme classifier on a large-scale speech recognition dataset, LibriSpeech, and use the gradient of the classifier during sampling. Although our unconditional generative

[*]Equal contribution [1]Data Science and AI Lab., Seoul National University [2]Department of ECE and Interdisciplinary Program in AI, Seoul National University. Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

model is trained without any transcript, Guided-TTS effectively generates mel-spectrograms given the transcript by guiding the generative process of unconditional DDPM using the phoneme classifier. As mispronunciation through guiding error is fatal for the TTS model, we present norm-based guidance that balances the classifier gradient and the unconditional score during sampling.

We demonstrate that Guided-TTS matches the performance of publicly available high-quality TTS models on LJSpeech without using LJSpeech transcripts. In addition, Guided-TTS generalizes well for diverse untranscribed datasets, and even for a long-form unsegmented dataset (Blizzard 2013). Furthermore, we show that the norm-based guidance significantly reduces pronunciation errors, which allows our proposed model to have a similar level of pronunciation accuracy as the existing conditional TTS models. We encourage readers to listen to samples of Guided-TTS trained on various untranscribed datasets on our demo page.[1]

2. Background

2.1. Denoising Diffusion Probablistic Models (DDPM) and Its Variant

DDPM (Sohl-Dickstein et al., 2015; Ho et al., 2020), which is proposed as a type of probabilistic generative model, has recently been applied to various domains, such as images (Dhariwal & Nichol, 2021) and audio (Chen et al., 2021a; Popov et al., 2021). DDPM first defines a forward process that gradually corrupts data X_0 into random noise X_T across T timesteps. The model learns the reverse process, which follows the reverse trajectory of the predefined forward process to generate data from random noise.

Recently, approaches have been proposed to formulate the trajectory between data and noise as a continuous stochastic differential equation (SDE) instead of using a discrete-time Markov process (Song et al., 2021b). Grad-TTS (Popov et al., 2021) introduces SDE formulation to TTS, which we have followed and used. According to the formulation of Grad-TTS, the forward process that corrupts data X_0 into standard Gaussian noise X_T is as follows:

$$dX_t = -\frac{1}{2}X_t\beta_t dt + \sqrt{\beta_t}dW_t, \tag{1}$$

where β_t is a predefined noise schedule, $\beta_t = \beta_0 + (\beta_T - \beta_0)t$, and W_t is a Wiener process. Anderson (1982) shows that the reverse process, which represents the trajectory from noise X_T to X_0, can also be formulated in SDE, which is defined as follows:

$$dX_t = (-\frac{1}{2}X_t - \nabla_{X_t}\log p_t(X_t))\beta_t dt + \sqrt{\beta_t}d\widetilde{W}_t, \tag{2}$$

where $\widetilde{W}_t$ is a reverse-time Wiener process. Given the score, the gradient of log density with respect to data (*i.e.,* $\nabla_{X_t}\log p_t(X_t)$), for $t \in [0, T]$, we can sample data X_0 from random noise X_T by solving Eq. (2). To generate data, the DDPM learns to estimate the score using the neural network s_θ parameterized by θ.

To estimate the score, X_t is sampled from the distribution derived from Eq. (1), given data X_0, which is as follows:

$$X_t|X_0 \sim \mathcal{N}(\rho(X_0, t), \lambda(t)), \tag{3}$$

where $\rho(X_0, t) = \mathrm{e}^{-\frac{1}{2}\int_0^t \beta_s ds}X_0$, and $\lambda(t) = I - \mathrm{e}^{-\int_0^t \beta_s ds}$. The score can then be derived from Eq. (3); $\nabla_{X_t}\log p_t(X_t|X_0) = -\lambda(t)^{-1}\epsilon_t$, where ϵ_t is the standard Gaussian noise used to sample X_t given X_0 (Popov et al., 2021). To train the model $s_\theta(X_t, t)$ for $\forall t \in [0, T]$, the following loss is used:

$$L(\theta) = \mathbb{E}_t\mathbb{E}_{X_0}\mathbb{E}_{\epsilon_t}\left[\left\|s_\theta(X_t, t) + \lambda(t)^{-1}\epsilon_t\right\|_2^2\right], \tag{4}$$

which is a L2 loss as in previous works (Ho et al., 2020; Song et al., 2021b).

Using model $s_\theta(X_t, t)$, we can generate sample X_0 from noise by solving Eq. (2). Grad-TTS generates data X_0 from X_T by setting $T = 1$ and using a fixed discretization strategy (Song et al., 2021b):

$$X_{t-\frac{1}{N}} = X_t + \frac{\beta_t}{N}(\frac{1}{2}X_t + \nabla_{X_t}\log p_t(X_t)) + \sqrt{\frac{\beta_t}{N}}z_t, \tag{5}$$

where N is the number of steps required to solve SDE, $t \in \{\frac{1}{N}, \frac{2}{N}, ..., 1\}$ and z_t is standard Gaussian noise.

2.2. Classifier Guidance

DDPM can be guided to generate samples with the desired condition without fine-tuning through the introduction of a classifier. Song et al. (2021b) use unconditional DDPM to generate class-conditional images by applying a separately trained image classifier. For conditional generation, the classifier $p_t(y|X_t)$ is trained to classify noisy data X_t as condition y.

Discretized SDE for conditional generation can be obtained by replacing the unconditional score $\nabla_{X_t}\log p_t(X_t)$ in Eq. (5) with a conditional score $\nabla_{X_t}\log p_t(X_t|y)$.

$$X_{t-\frac{1}{N}} = X_t + \frac{\beta_t}{N}(\frac{1}{2}X_t + \nabla_{X_t}\log p_t(X_t|y)) + \sqrt{\frac{\beta_t}{N}}z_t, \tag{6}$$

$$\nabla_{X_t}\log p_t(X_t|y) = \nabla_{X_t}\log p_t(X_t) + \nabla_{X_t}\log p_t(y|X_t). \tag{7}$$

If the unconditional score and classifier gradient for the target condition are given, the sample X_0 with condition y can be generated using Eq. (6).

[1] Demo : https://bit.ly/3r8vho7

Dhariwal & Nichol (2021) guide not only unconditional DDPM but also conditional DDPM using a classifier. They introduce a gradient scale s when guiding the DDPM, which is multiplied by the classifier gradient ($s \cdot \nabla_{X_t} \log p_t(y|X_t)$) to adjust the scale of it. By using $s > 1$, they generate higher-fidelity (but less diverse) samples, which contributes to achieving the state-of-the-art performance for class-conditional image generation.

3. Guided-TTS

由四部分组成：无条件DDPM、音素分类器、持续时间预测器和说话者编码器

In this section, we present Guided-TTS, which aims to build a high-quality text-to-speech model without any transcript of the target speaker. Whereas other TTS models directly learn to generate speech from text, Guided-TTS learns to model unconditional distribution of speech and generates speech from text using classifier guidance. For classifier guidance, we train an unconditional diffusion model on untranscribed speech data and leverage a phoneme classifier trained on a large-scale speech recognition dataset. To the best of our knowledge, Guided-TTS is the first TTS model to generate speech using an unconditional generative model.

Guided-TTS consists of four modules: unconditional DDPM, phoneme classifier, duration predictor, and speaker encoder, as shown in Fig. 1. The unconditional DDPM learns to generate mel-spectrogram unconditionally, and the remaining three modules are used for TTS synthesis through guidance. We describe the unconditional DDPM in Section 3.1, followed by the method of guiding the unconditional model for TTS in Section 3.2.

3.1. Unconditional DDPM

Our unconditional DDPM models the unconditional distribution of speech P_X without any transcript. We use untranscribed speech data from a single target speaker S as the training data for the diffusion model to build a TTS for the speaker S. Since our diffusion model learns without transcript, training samples do not need to be aligned with the transcripts. Thus, we use random chunks of untranscribed speech data as training data such that Guided-TTS does not require speech transcription and sentence-level segmentation when only the long-form untranscribed data is available for the target speaker S.

Given a mel-spectrogram $X = X_0$, we define the forward process as in Eq. (1), which gradually corrupts data into noise, and approximate the reverse process in Eq. (2), by estimating the unconditional score $\nabla_{X_t} \log p(X_t)$ for each timestep t. At each iteration, $X_t, t \in [0, 1]$ is sampled from the mel-spectrogram X_0 as in Eq. (3), and the score is estimated using the neural network $s_\theta(X_t, t)$ parameterized by θ. The training objective of the unconditional model is given by Eq. (4).

Similar to Grad-TTS (Popov et al., 2021), we regard mel-spectrogram as a 2D image with a single channel and use the U-Net architecture (Ronneberger et al., 2015) as s_θ. We use the same sized architecture applied to model 32×32 sized images in Ho et al. (2020) to capture long-term dependencies without any text information, whereas Grad-TTS uses a smaller architecture for the conditional distribution modeling.

3.2. Text-to-Speech via Classifier Guidance

For TTS synthesis, we introduce a frame-wise phoneme classifier and use a classifier guidance method to guide unconditional DDPM. TTS via classifier guidance decouples the generative modeling of speech by conditioning text information. This allows us to leverage a noisy speech recognition dataset as training data for the phoneme classifier, which is challenging to utilize for training existing TTS models.

As shown in Fig. 1, in order to generate mel-spectrogram given text, our duration predictor outputs the duration for each text token and expands the transcript y to frame-level phoneme label $\hat{y}$. We then sample a random noise X_T of the same length as $\hat{y}$ from the standard normal distribution, and we can generate conditional samples with a conditional score. As in Eq. (8), we can estimate the conditional score on the left side by adding the two terms on the right side: the first term is obtained from the unconditional DDPM, and the second term can be computed using the phoneme classifier. That is, we build a text-to-speech model with the unconditional generative model for speech by adding the gradient of the phoneme classifier during the generative process.

$$\nabla_{X_t} \log p(X_t|\hat{y}, spk = S) = \nabla_{X_t} \log p_\theta(X_t|spk = S) + \nabla_{X_t} \log p_\phi(\hat{y}|X_t, spk = S) \tag{8}$$

To guide the unconditional DDPM for any target speaker S, our phoneme classifier and duration predictor are trained on a large-scale speech recognition dataset and designed to be speaker-dependent modules for better generalization to the unseen speaker S. We provide the speaker embedding extracted from the pre-trained speaker verification network as a condition for both modules, as described in Fig. 1. We describe each module required for guidance below.

Phoneme Classifier The phoneme classifier is a network trained on a large-scale speech recognition dataset that recognizes the phoneme corresponding to each frame of the input mel-spectrogram. To train the frame-wise phoneme classifier, we align transcript and speech using a forced alignment tool, the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), and extract the frame-level phoneme label $\hat{y}$. The phoneme classifier is trained to classify the corrupted

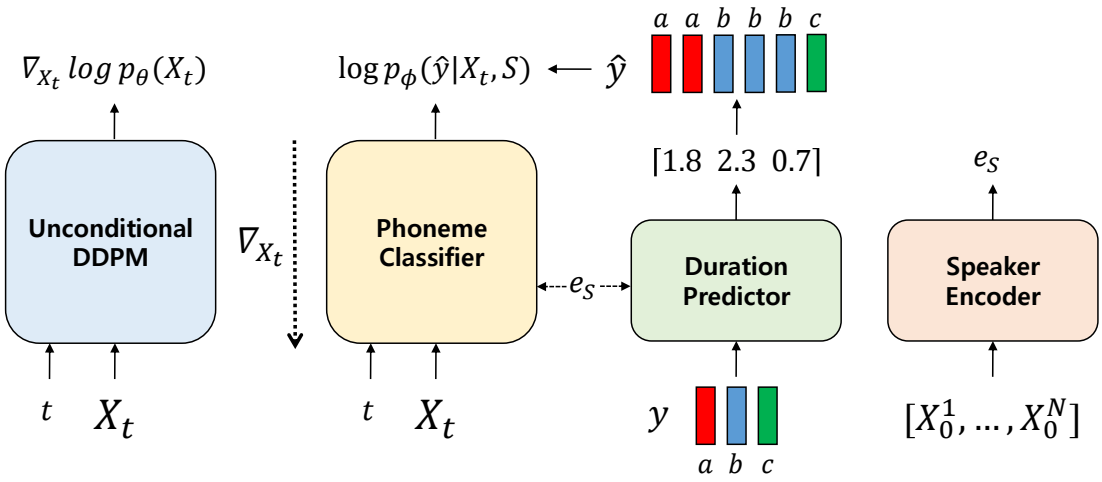

Figure 1: The overall architecture of Guided-TTS. The unconditional DDPM learns to generate speech X_0 with untranscribed data. The other modules, the phoneme classifier, duration predictor, and speaker encoder are for guiding the unconditional DDPM to generate conditional samples given y.

mel-spectrogram X_t sampled from Eq. (3) as a frame-level phoneme label $\hat{y}$. The training objective of the phoneme classifier is to minimize the expectation of cross-entropy between the phoneme label $\hat{y}$ and output probability with respect to $t \in [0, 1]$.

We use a WaveNet-like architecture (van den Oord et al., 2016) as a phoneme classifier, and time embedding e_t, which is extracted in the same way as in Popov et al. (2021), is used as a global condition in WaveNet to provide information regarding the noise level of the corrupted input X_t at timestep t. For speaker-dependent classification, we also use the speaker embedding e_S from the speaker encoder as the global condition.

Duration Predictor The duration predictor is a module that predicts the duration of each text token for a given text sequence y. We extract the duration label of each text token using MFA for the same data on which the phoneme classifier is trained. The duration predictor is trained to minimize L2 loss between the duration label and the estimated duration in the log-domain, and we round up the estimated duration during inference. The architecture of the duration predictor is the same as that of Glow-TTS (Kim et al., 2020) with the text encoder. We concatenate the text embedding and speaker embedding e_S to predict the speaker-dependent duration.

Speaker Encoder The speaker encoder encodes the speaker information from the input mel-spectrogram and outputs the speaker embedding e_S. Similar to Jia et al. (2018), we train a speaker encoder with GE2E loss (Wan et al., 2018) on the speaker verification dataset and use the speaker encoder to condition speaker-dependent modules. We extract the speaker embedding e_S from the clean mel-spectrogram X_0 for each training data. For guidance, we average and normalize the speaker embeddings of the untranscribed speech for the target speaker S to extract e_S.

3.2.1. Norm-based Guidance

Algorithm 1 Norm-based Guidance

$\hat{y}$: frame-wise phoneme label, s: gradient scale, τ: temperature
θ: parameter of unconditional DDPM:, ϕ: parameter of phoneme classifier
$X_1 \sim \mathcal{N}(0, \tau^{-1}I)$
for $i = N$ **to** 1 **do**
  $t \leftarrow \frac{i}{N}$
  $\alpha_t \leftarrow \|\nabla_{X_t} \log p_\theta(X_t)\| \,/\, \|\nabla_{X_t} \log p_\phi(\hat{y}|X_t)\|$
  $z_t \sim \mathcal{N}(0, \tau^{-1}I)$
  $\mu_t \leftarrow \frac{1}{2}X_t + \nabla_{X_t} \log p_\theta(X_t) + s \cdot \alpha_t \nabla_{X_t} \log p_\phi(\hat{y}|X_t)$
  $X_{t-\frac{1}{N}} \leftarrow X_t + \frac{\beta_t}{N}\mu_t + \sqrt{\frac{\beta_t}{N}} z_t$
end for
return X_0

Initially, we scaled the gradient of the classifier $\nabla_{X_t} \log p_\phi(\hat{y}|X_t, spk = S)$ in Eq. (8) using gradient scale s (Dhariwal & Nichol, 2021). However, when guiding the unconditional DDPM with the frame-wise phoneme classifier, we found that the norm of the unconditional score suddenly increases near $t = 0$ (Appendix A.4). That is, when closer to data X_0, the phoneme classifier has little effect on the generative process of the DDPM. As a matter of fact, our experiments on generating samples using various numbers of gradient scale s resulted in mispronouncing samples given text for all cases.

Herein, we propose norm-based guidance to guide the unconditional DDPM better in terms of generating speech conditioned on frame-level phoneme label $\hat{y}$. Norm-based guidance is a method of scaling the norm of the classifier gradient in proportion to the norm of the score in order to prevent the effect of the gradient from being insignificant as the score steeply increases. The ratio between the norm of the scaled gradient and the norm of the score is defined as the gradient scale s. By adjusting s, we can determine how much the classifier gradient contributes to the guidance of unconditional DDPM. We also use the temperature parameter τ when guiding the DDPM. We observe that tuning τ to a value greater than 1 helps generate high-quality mel-spectrograms. Detailed analysis on classifier guidance are in section 5.3.

4. Experiments

数据集

Datasets In Guided-TTS, the speaker-dependent phoneme classifier and duration predictor are trained on LibriSpeech (Panayotov et al., 2015), which is a large-scale automatic speech recognition (ASR) dataset with approximately 982 hours of speech uttered by 2,484 speakers with corresponding texts. To extract the speaker embedding e_S from each utterance, we train a speaker encoder on VoxCeleb2 (Chung et al., 2018), which is a speaker verification dataset that contains more than 1M utterances of 6112 speakers.

For the comparison case with baselines which make use of the target speaker transcript data, we use LJSpeech (Ito, 2017), a 24-hour female single speaker dataset consisting of 13,100 audio clips. For the other case which makes use of only the untranscribed target speaker speech, we use LJSpeech, Hi-Fi TTS (Bakhturina et al., 2021), and Blizzard 2013 (King & Karaiskos, 2013). Hi-Fi TTS is a multi-speaker TTS dataset with 6 females and 4 males, and the data of each speaker consists of at least 17 hours of speech. We select three relatively clean speakers among them (two males (ID: 6097, 9017) and one female (ID: 92)). Blizzard 2013 is a 147 hours-long audiobook containing both segmented and unsegmented data read by a single female speaker. We use the unsegmented data of Blizzard 2013, randomly clipping 5-seconds-long chunks of audio to build a TTS for long-form untranscribed data.

Training Details We convert text into International Phonetic Alphabet (IPA) phoneme sequences using open-source software (Bernard, 2021). To extract the mel-spectrogram, we use the same hyperparameters as Glow-TTS (Kim et al., 2020). All modules are trained using Adam optimizer with a learning rate of 0.0001. For the unconditional model and the phoneme classifier, $\beta_0 = 0.05$ and $\beta_1 = 20$ are used for beta schedule. Other details and hyperparameters of Guided-TTS are described in Appendix A.1.

Evaluation To compare the performance of models with transcribed data, we use the official implementations and pre-trained models of Glow-TTS and Grad-TTS.[2][3] For Glow-TTS, we use a pre-trained model with blank tokens between phonemes and use $\tau = 1.5$. We use the same hyperparameters as the official implementation, $\tau = 1.5$, and the number of reverse steps $N = 50$ for Grad-TTS. To compare model performance in the absence of a transcript, we extract the transcript using a CTC-based conformer-large ASR model (Graves et al., 2006; Gulati et al., 2020) from NEMO toolkit (Kuchaiev et al., 2019), which is pre-trained using LibriSpeech. We train Grad-TTS using the ASR transcribed data for 1.7m iterations, which we refer to as Grad-TTS-ASR. For Guided-TTS, we set $\tau = 1.5$, and the number of reverse steps $N = 50$. We observe that the low classification accuracy of the phoneme classifier near $t = 1$ (closer to random noise) deteriorates the sample quality. Therefore, we refrain from using the gradient of the classifier at the initial steps of sampling, setting the gradient scale s to 0. Afterwards, we linearly increase the gradient scale s to 0.3. For the vocoder, we use the official implementation and pre-trained models of HiFi-GAN.[4]

To show whether Guided-TTS with norm-based guidance generates the sentences of the given text accurately, we measure the character error rate (CER) for each model, which is a metric commonly used in automatic speech recognition (ASR). To compute the metric, we use the CTC-based conformer pre-trained with 7,000 hours of speech from the NEMO toolkit. We generate 5 samples for each sentence in the test set and measure all CER for all generated samples, ultimately using the whole average CER for comparison.

5. Results

5.1. Model Comparison

We compare the performances of audio samples by measuring the 5-scale mean opinion score (MOS) on LJSpeech using Amazon Mechanical Turk. In addition, through CER, we check whether the generated sample of each model faithfully reflects the text. To calculate the CER, we first synthesize the speech of a given text for each model and provide it to the ASR model to extract the text corresponding to the generated sample. We then measure the CER between the ground truth text and the text obtained from the ASR model. For evaluation, we randomly select 50 samples drawn from the test set of LJSpeech and measure the MOS and CER.

In Table 1, we compare the performance and CER of Guided-TTS with Glow-TTS and Grad-TTS, which are high-quality TTS models. While Glow-TTS and Grad-TTS use tran-

[2]Glow-TTS: https://bit.ly/3kS315K
[3]Grad-TTS: https://bit.ly/3qTCmcJ
[4]HiFi-GAN: https://bit.ly/3FxBv5x

Table 1: Mean Opinion Score (MOS) with 95% confidence intervals of TTS models for LJSpeech. "GT MEL" represents the HiFi-GAN result of ground truth mel-spectrogram.

Method	LJ Transcript	5-scale MOS	CER(%)
GT		4.45 ± 0.05	0.64
GT MEL		4.24 ± 0.07	0.77
GLOW-TTS	$\surd$	4.14 ± 0.08	0.66
GRAD-TTS	$\surd$	4.25 ± 0.07	1.09
GUIDED-TTS	$\times$	4.25 ± 0.08	1.03

scribed data of LJSpeech, Guided-TTS only uses untranscribed data of LJSpeech to train unconditional DDPM. Guided-TTS shows comparable performance to other TTS models without any transcript of LJSpeech by leveraging the phoneme classifier trained on LibriSpeech. Guided-TTS also has a similar CER to that of the conditional TTS models, which shows that the unconditional DDPM accurately generates speech from the given transcripts using norm-based classifier guidance. This demonstrates that our proposed model enables the building of a high-quality TTS model without any transcript of the target speaker. Samples of all models are available on the demo page.[5]

5.2. Generalization to Diverse Datasets

In the previous section, we showed that Guided-TTS can synthesize high-quality speech without transcript of LJSpeech. Since we separate the training of the unconditional model and the classifier, we are capable of building TTS models for various untranscribed datasets by combining the single phoneme classifier to various unconditional DDPMs trained on untranscribed datasets.

In this section, we assume that only untranscribed speech is available for each speaker. Since existing TTS models inevitably require data with transcripts for training, we extract transcripts from the various untranscribed datasets using a pre-trained ASR model in order to train the powerful baseline, Grad-TTS. We refer to this baseline as Grad-TTS-ASR. Since Guided-TTS leverages the phoneme classifier trained on LibriSpeech, the specific ASR model pre-trained on LibriSpeech is selected to extract transcriptions, making fair comparison possible. For various datasets, we compare Guided-TTS with Grad-TTS-ASR. We use 50 randomly chosen sentences from the test set of each dataset.

The performance of each model on LJSpeech and Hi-Fi TTS is presented in Table 2. For LJSpeech, both Guided-TTS and Grad-TTS-ASR achieve comparable performances to Grad-TTS using transcript. However, for Hi-Fi TTS, Guided-TTS outperforms Grad-TTS-ASR and exhibits low CER values for all datasets. This shows that the single phoneme classifier

[5]Demo : https://bit.ly/3r8vho7

Table 2: Mean Opinion Score (MOS) with 95% confidence intervals of TTS models for multiple datasets. "Data" refers to the untranscribed speech dataset used for each model. For Blizzard, we use long-form unsegmented data for training.

Data	Method	5-scale MOS	CER(%)
LJSPEECH	GT	4.45±0.05	0.64
	GT MEL	4.24±0.07	0.77
	GRAD-TTS	4.25±0.07	1.09
	GRAD-TTS-ASR	4.23±0.08	1.16
	GUIDED-TTS	4.25±0.08	1.03
HI-FI TTS (ID: 92)	GT	4.48±0.07	0.09
	GT MEL	4.27±0.07	0.20
	GRAD-TTS-ASR	4.11±0.08	1.33
	GUIDED-TTS	4.20±0.08	0.81
HI-FI TTS (ID: 6097)	GT	4.50±0.05	0.24
	GT MEL	4.26±0.07	0.33
	GRAD-TTS-ASR	4.09±0.08	1.88
	GUIDED-TTS	4.16±0.08	0.79
HI-FI TTS (ID: 9017)	GT	4.45±0.05	0.11
	GT MEL	4.21±0.07	0.07
	GRAD-TTS-ASR	3.83±0.09	2.04
	GUIDED-TTS	4.04±0.09	0.21
BLIZZARD	GT	4.44±0.05	0.51
	GT MEL	4.26±0.09	0.48
	GUIDED-TTS	4.24±0.09	0.24

of Guided-TTS stably generates the given text for various datasets. On the other hand, we confirm that the pronunciation accuracy and sample quality of Grad-TTS-ASR, which uses the noisy transcript generated by ASR, are not robust to dataset. Aside from this, we demonstrate that Guided-TTS can robustly generate out-of-distribution (OoD) text for several datasets. The results and details of generated OoD texts are provided in Appendix A.3. We also show the performance of Guided-TTS trained with random chunks of unsegmented data of Blizzard 2013, a long-form audiobook dataset, in Table 2. Guided-TTS generates high-quality samples without the transcript of Blizzard dataset, just like it has done with the other datasets. In addition, the low CER of Guided-TTS indicates that a TTS model of accurate pronunciation can be built even when using randomly cropped audio without sentence-level segmentation for training.

Based on the results above, we demonstrate that the proposed method enables TTS for untranscribed datasets of various characteristics (*e.g.,* gender, accent, and prosody). Samples on various speakers are available on the demo page.

5.3. Analysis

Norm-based Guidance We also compare the proposed norm-based classifier guidance with the classifier guidance used in previous works (Song et al., 2021b; Dhariwal & Nichol, 2021). A model that conducts a conditional gen-

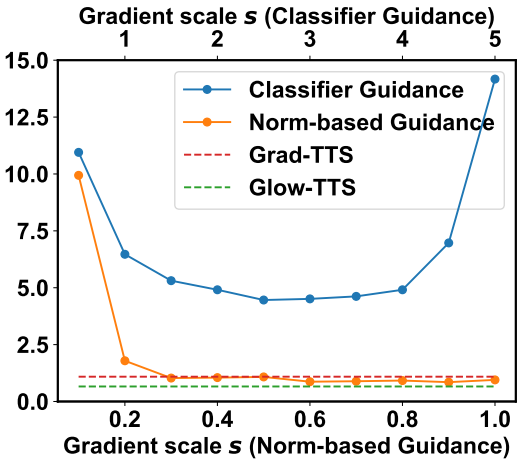

Figure 2: CER of Guided-TTS with classifier guidance (Dhariwal & Nichol, 2021) and norm-based guidance according to gradient scales.

eration task with classifier guidance occasionally generates samples of conditions other than the target condition (Song et al., 2021b). Similarly, we observe that Guided-TTS with the classifier guidance method produces mispronounced samples given text. To show the effect of norm-based guidance and adjustment of the gradient scale, we measure the CER of Guided-TTS for LJSpeech according to the gradient scale s. We explore the gradient scale s within $[0.5, 1.0, ..., 5.0]$ for classifier guidance (Dhariwal & Nichol, 2021), and $[0.1, 0.2, ..., 1.0]$ for norm-based guidance.

Fig. 2 presents the CER of Guided-TTS with the classifier guidance (Dhariwal & Nichol, 2021) and the proposed norm-based classifier guidance. As shown in Fig. 2, the sample generated using the existing guidance method shows a far worse CER than the existing TTS models, which indicates that it is unsuitable for TTS. By contrast, the proposed guidance method with the appropriate gradient scale helps accurately generate samples given text sentences, similar to existing TTS models.

If the gradient scale is too small, the effect of the classifier gradient is negligible, and the generated samples do not reflect the given text. On the other hand, we observed that guidance with a large gradient scale deteriorates the sample quality. For the proposed norm-based guidance, we set the default gradient scale s to 0.3, which generates high-quality samples that exactly match the given text. Samples for multiple gradient scales with each guidance method are on the demo page.

Amount of Data for Phoneme Classifier We show the CER of Guided-TTS on LJSpeech according to the amount of LibriSpeech data used for training the phoneme classifier in Table 3. We train the phoneme classifiers with 1% (9 hours), 10% (96 hours), 100% (960 hours) of LibriSpeech

Table 3: CER of Guided-TTS on LJSpeech test set according to the amount of data used for training the phoneme classifier.

Method	CER(%)
GUIDED-TTS (LIBRISPEECH 100%)	1.03
GUIDED-TTS (LIBRISPEECH 10%)	2.28
GUIDED-TTS (LIBRISPEECH 1%)	4.24

respectively, and the classification accuracy of each model is shown in Fig. 3. The CER results in Table 3 indicate that the amount of data used for the phoneme classifier is critical regarding the pronunciation accuracy of Guided-TTS. Therefore, the pronunciation of Guided-TTS improves as the amount of data used for phoneme classification increases. Thus, we anticipate that Guided-TTS can be improved even further with a much larger-scale ASR dataset.

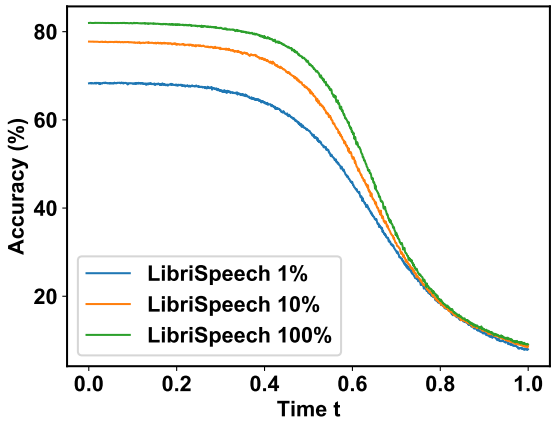

Figure 3: Phoneme classification accuracy on LJSpeech test set over timestep t. The number next to the term LibriSpeech indicates the portion of LibriSpeech used for training.

6. Related Work

Unconditional Speech Generation In general, the unconditional speech generative model (van den Oord et al., 2016; Vasquez & Lewis, 2019), which models audio without any information, is more challenging than the conditional generative model that synthesizes speech using text or mel-spectrograms. Several works have attempted to unconditionally generate raw waveforms (van den Oord et al., 2016; Donahue et al., 2019) or to model the unconditional distribution of latent code or mel-spectrogram of audio (van den Oord et al., 2017; Vasquez & Lewis, 2019; Lakhotia et al., 2021; Kharitonov et al., 2022) instead of directly modeling raw waveforms. Most existing unconditional models have only been used for unconditional audio modeling and no other purposes. To the best of our knowledge, this is the first application of an unconditional model for TTS with

appropriate guidance to enable speech synthesis using untranscribed data from a target speaker.

Text-to-Speech Models Most text-to-speech (TTS) models are composed of two parts: a model that generates intermediate features (*e.g.,* mel-spectrogram) from text (Shen et al., 2018) and a vocoder, which synthesizes raw waveforms from intermediate features (van den Oord et al., 2016). The autoregressive model is used for the text-to-intermediate feature model (Wang et al., 2017; Shen et al., 2018; Ping et al., 2018; Li et al., 2019) and vocoder (van den Oord et al., 2016; Kalchbrenner et al., 2018) to perform high-quality TTS. To improve the sampling speed of the autoregressive models, flow-based generative models (Kingma & Dhariwal, 2018) and feed-forward models have been proposed for text-to-mel-spectrogram models (Ren et al., 2019; 2021; Kim et al., 2020; Shih et al., 2021) and vocoders (Oord et al., 2018; Prenger et al., 2019; Kim et al., 2019). In addition, variational autoencoder based models (Kingma & Welling, 2014; Lee et al., 2020; Liu et al., 2021), diffusion based models (Ho et al., 2020; Chen et al., 2021a; Kong et al., 2021; Popov et al., 2021; Jeong et al., 2021), and GAN based models (Goodfellow et al., 2014; Kumar et al., 2019; Bińkowski et al., 2019; Kong et al., 2020) have been proposed as high-quality speech synthesis models with parallel sampling schemes. End-to-end TTS models have recently been proposed, such as Ren et al. (2021), Donahue et al. (2021), Weiss et al. (2021), Kim et al. (2021), and Chen et al. (2021b).

Most previous TTS models perform conditional generation tasks using transcribed data of the target speaker. On the other hand, Guided-TTS models unconditional distribution of speech with untranscribed data and generates conditional samples with the pre-trained phoneme classifier. By modeling unconditional distribution of speech, Guided-TTS can utilize long-form untranscribed data of the target speaker without sentence-level segmentation or transcription.

Text-to-Speech with Untranscribed Data There are two main approaches when building a TTS model without the target speaker's transcript: fine-tuning based approach and speaker embedding based approach. Both approaches require a pre-trained multi-speaker TTS model. In the fine-tuning based approach (Yan et al., 2021), the mel-spectrogram encoder is combined with the pre-trained TTS model to fine-tune the model with untranscribed speech of the target speaker. Speaker embedding based approach (Arik et al., 2018; Jia et al., 2018; Casanova et al., 2021) provides the target speaker's embedding extracted from untranscribed speech to the TTS model for adaptation. These methods require a large-scale multi-speaker TTS dataset, which is difficult to collect and challenging to model the distribution. Also, the performances of these approaches are worse than single speaker TTS models (Kim et al., 2020; Ren et al., 2021; Popov et al., 2021) trained with the <speech, text> pair of the target speaker. On the other hand, instead of using a multi-speaker TTS dataset, we utilize an automatic speech recognition (ASR) dataset to build a TTS model, which is relatively easy to collect. By leveraging the phoneme classifier trained on the ASR dataset, Guided-TTS achieves performance comparable to other TTS models (Kim et al., 2020; Ren et al., 2021; Popov et al., 2021) with untranscribed data of the target speaker.

There is also an approach that utilizes an untranscribed dataset to extract unsupervised linguistic units and reduces the amount of the paired dataset (Zhang & Lin, 2020). This model focuses on TTS for low-resource languages, while Guided-TTS assumes that a large-scale speech recognition dataset is available and only untranscribed data is given for the target speaker.

Diffusion-based Generative Models DDPM (Sohl-Dickstein et al., 2015; Ho et al., 2020) has undergone several theoretical developments (Song et al., 2021b) and produces high quality samples in many domains (Ho et al., 2020; Dhariwal & Nichol, 2021; Chen et al., 2021a; Popov et al., 2021; Luo & Hu, 2021). A continuous version of DDPM, an SDE-based model (Song et al., 2021b; Popov et al., 2021) is also presented. Thanks to many theoretical and practical breakthroughs (Song et al., 2021a; Nichol & Dhariwal, 2021), DDPM has also shown strong performance in speech synthesis (Chen et al., 2021a; Kong et al., 2021; Popov et al., 2021; Jeong et al., 2021).

A pre-trained unconditional DDPM can be used for various tasks such as imputation (Song et al., 2021b), and controllable generation (Song et al., 2021b). In particular, the controllable generation allows (Dhariwal & Nichol, 2021) to achieve state-of-the-art performance in class-conditional image generation by guiding the DDPM using a gradient from the classifier trained on the same dataset as DDPM. We introduce the classifier guidance method of unconditional DDPM to text-to-speech synthesis. Our unconditional DDPM and the phoneme classifier can be trained using different datasets, making it possible to build a TTS model with the target speaker's untranscribed speech.

7. Conclusion

In this work, we present Guided-TTS, a new type of TTS model that generates speech given transcript by guiding the unconditional diffusion-based model for speech. As Guided-TTS models unconditional distribution for speech, we can construct a TTS model using the target speaker's untranscribed data. Thanks to the properties of diffusion-based generative models, our unconditional generative model can generate a speech when a transcript is given by introducing the phoneme classifier trained on LibriSpeech. To the

best of our knowledge, Guided-TTS is the first TTS model to leverage the unconditional generative model for speech. We showed that Guided-TTS matches the performance of the previous TTS models on LJSpeech without the transcript. We also showed that Guided-TTS generalizes well to diverse untranscribed datasets with the single phoneme classifier. We believe that Guided-TTS can reduce the burden of constructing training datasets for high-quality TTS.

Acknowledgements

We would like to thank Jiheum Yeom and Jaehyeon Kim for their helpful discussions. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00959), AIRS Company in Hyundai Motor and Kia through HMC/KIA-SNU AI Consortium Fund, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022.

References

Anderson, B. D. O. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3):313–326, May 1982.

Arik, S., Chen, J., Peng, K., Ping, W., and Zhou, Y. Neural voice cloning with a few samples. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/4559912e7a94a9c32b09d894f2bc3c82-Paper.pdf.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., and Zhang, Y. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*, 2021.

Bernard, M. Phonemizer. https://github.com/bootphon/phonemizer, 2021.

Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K. High Fidelity Speech Synthesis with Adversarial Networks. In *International Conference on Learning Representations*, 2019.

Casanova, E., Shulby, C., Gölge, E., Müller, N. M., de Oliveira, F. S., Candido Jr., A., da Silva Soares, A., Aluisio, S. M., and Ponti, M. A. SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model. In *Proc. Interspeech 2021*, pp. 3645–3649, 2021. doi: 10.21437/Interspeech.2021-1774.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*, 2021a.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., Dehak, N., and Chan, W. WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis. In *Proc. Interspeech 2021*, pp. 3765–3769, 2021b. doi: 10.21437/Interspeech.2021-1897.

Chung, J. S., Nagrani, A., and Zisserman, A. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Donahue, C., McAuley, J., and Puckette, M. Adversarial audio synthesis. *International Conference on Learning Representations (ICLR)*, 2019.

Donahue, J., Dieleman, S., Binkowski, M., Elsen, E., and Simonyan, K. End-to-end Adversarial Text-to-Speech. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rsf1z-JSj87.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33. Curran Associates, Inc., 2020.

Ito, K. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Jeong, M., Kim, H., Cheon, S. J., Choi, B. J., and Kim, N. S. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. In *Proc. Interspeech 2021*, pp. 3605–3609, 2021. doi: 10.21437/Interspeech.2021-469.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, z., Nguyen, P., Pang, R., Lopez Moreno, I., and Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.

Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T. A., Riviere, M., Mohamed, A., Dupoux, E., and Hsu, W.-N. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.593. URL https://aclanthology.org/2022.acl-long.593.

Kim, J., Kim, S., Kong, J., and Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Advances in Neural Information Processing Systems*, 33, 2020.

Kim, J., Kong, J., and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv:2106.06103*, 2021.

Kim, S., Lee, S.-G., Song, J., Kim, J., and Yoon, S. Flowavenet: A generative flow for raw audio. In *International Conference on Machine Learning*, pp. 3370–3378, 2019.

King, S. J. and Karaiskos, V. The blizzard challenge 2013. In *In Blizzard Challenge Workshop*, 2013.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10236–10245, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Kong, J., Kim, J., and Bae, J. HiFi-GAN: Generative Adversarial networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, 33, 2020.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*, 2021.

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. C. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems 32*, pp. 14910–14921, 2019.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., and Dupoux, E. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00430. URL https://doi.org/10.1162/tacl_a_00430.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Lee, Y., Shin, J., and Jung, K. Bidirectional variational inference for non-autoregressive text-to-speech. In *International Conference on Learning Representations*, 2020.

Li, N., Liu, S., Liu, Y., Zhao, S., and Liu, M. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6706–6713, 2019.

Liu, P., Cao, Y., Liu, S., Hu, N., Li, G., Weng, C., and Su, D. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. *arXiv preprint arXiv:2102.06431*, 2021.

Luo, S. and Hu, W. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2837–2845, June 2021.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*, 2017.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Ping, W., Peng, K., Gibiansky, A., Arik, S. Ö., Kannan, A., Narang, S., Raiman, J., and Miller, J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *International Conference on Learning Representations*, 2018.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8599–8608. PMLR, 2021.

Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621. IEEE, 2019.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech: Fast, Robust and Controllable Text to Speech. volume 32, pp. 3171–3180, 2019.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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

Shih, K. J., Valle, R., Badlani, R., Lancucki, A., Ping, W., and Catanzaro, B. Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

van den Oord, A., Vinyals, O., et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

Vasquez, S. and Lewis, M. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*, 2019.

Wan, L., Wang, Q., Papir, A., and Moreno, I. L. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883, 2018. doi: 10.1109/ICASSP.2018.8462665.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pp. 4006–4010, 2017. doi: 10.21437/Interspeech.2017-1452.

Weiss, R. J., Skerry-Ryan, R., Battenberg, E., Mariooryad, S., and Kingma, D. P. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5679–5683. IEEE, 2021.

Yan, Y., Tan, X., Li, B., Qin, T., Zhao, S., Shen, Y., and Liu, T.-Y. Adaspeech 2: Adaptive text to speech with untranscribed data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6613–6617. IEEE, 2021.

Zhang, H. and Lin, Y. Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. In Meng, H., 0011, B. X., and Zheng, T. F. (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 3161–3165. ISCA, 2020. doi: 10.21437/Interspeech.2020-1403. URL https://doi.org/10.21437/Interspeech.2020-1403.

A. Appendix

A.1. Training Details and Hyperparamters

In this section, we cover the training details and detailed hyperparameters of Guided-TTS. We only use untranscribed data of the various target speakers (LJSpeech, Hi-Fi TTS, and Blizzard 2013) for training unconditional DDPMs and we train the phoneme classifier and duration predictor on LibriSpeech. Alignment labels are required for training the phoneme classifier and the duration predictor, and we train Montreal Forced Aligner (MFA) on LibriSpeech to extract the alignment.

The unconditional DDPMs are trained with batch size 16 for all datasets. The phoneme classifier of Guided-TTS uses a WaveNet-like structure with 256 residual channels and 6 residual blocks stacks of 3 dilated convolution layers, and is trained for 200 epochs with batch size 64. The duration predictor is trained for 20 epochs with batch size 64. The speaker encoder is a two-layer LSTM with 768 channels followed by a linear projection layer to extract 256-dimensional speaker embedding e_S, and trained for 300K iterations.

For sampling, we use the last checkpoint for the unconditional DDPM and the speaker encoder. For the phoneme classifier and the duration predictor, we use the checkpoint of the epoch that scores best on its respective metric (validation accuracy for the phoneme classifier and validation loss of the duration predictor).

A.2. Hardware and Sampling Speed 硬件要求

We conduct all experiments and evaluations using NVIDIA's RTX A40 with 48GB memory. Although the main objectives of Guided-TTS are not focused on fast inference, it can perform real-time speech synthesis on GPU for $N = 50$, which is the number of reverse steps we use for evaluation. We measure the sampling speed of Guided-TTS using a real-time factor (RTF). We also measure how much time it takes to compute the unconditional score ($\nabla_{X_t} \log p_\theta(X_t)$) and gradient of the classifier ($\nabla_{X_t} \log p_\phi(\hat{y}|X_t)$). Guided-TTS achieves an RTF of 0.486, of which 0.184 is used to calculate the score and 0.291 is used for classifier gradient calculation.

A.3. Out-of-Distribution (OoD) Text Robustness

From section 5.2, we have confirmed that Guided-TTS constructs high-quality TTS models of various speakers. Leveraging the phoneme classifier well-trained on ASR data, Guided-TTS generates high-quality samples with precise pronunciation. Since the phoneme classifier is trained on large-scale ASR data, Guided-TTS generates samples from OoD texts robustly, those of which the model has not seen in the target speaker datasets. In Table 4, we show the performance and CER of the OoD samples generated by the Guided-TTS model trained on Hi-Fi TTS. We randomly select 50 sentences from LJSpeech's test set for the OoD text. And for comparison, we use Grad-TTS-ASR trained on Hi-Fi TTS to generate speech corresponding to the OoD text.

Table 4: Mean Opinion Score (MOS) with 95% confidence intervals of TTS models for out-of-distribution (OoD) text (LJSpeech test set). "Data" refers to the untranscribed speech dataset used for each model.

Data	Method	5-scale MOS	CER(%)
HI-FI TTS (ID: 92)	GRAD-TTS-ASR	4.14±0.08	2.15
	GUIDED-TTS	4.23±0.07	0.94
HI-FI TTS (ID: 6097)	GRAD-TTS-ASR	3.99±0.08	2.49
	GUIDED-TTS	4.18±0.08	0.97
HI-FI TTS (ID: 9017)	GRAD-TTS-ASR	3.91±0.09	2.74
	GUIDED-TTS	4.15±0.08	0.84

Through Table 4, we observe that Guided-TTS produces high-quality samples. The CER result of Table 4 indicates that Guided-TTS generates a sample that faithfully reflects the OoD text. On the other hand, we confirm that Grad-TTS-ASR produces inaccurate samples, showing worse quality for OoD text. Through these results, we demonstrate that Guided-TTS is a TTS model that robustly generates samples for diverse text.

A.4. Norm of the Unconditional Score and Classifier Gradient

The norm of the unconditional score and the gradient norm of the classifier for each timestep are shown in Fig. 4. We sample X_t at a total of 1000 timesteps ($t \in (\frac{1}{2000}, \frac{3}{2000}, ..., \frac{1999}{2000})$) using Eq. (3) for all 500 samples from the test set of the LJSpeech. We then obtain the norm of the unconditional score and the gradient of the classifier using the sampled X_t with Guided-TTS trained on LJSpeech. Each norm is averaged over 500 samples for each timestep. As shown in Fig. 4, the norm of the unconditional score rises steeply around $t = 0$. This is about 70 times larger than the norm of the classifier gradient near $t = 0$, which significantly reduces the effect of the classifier guidance. To alleviate this problem, we propose the norm-based guidance in Section 3.2.1, which helps prevent both the gradient of the classifier from being ignored and the issue of synthesized speech not matching the text.

A.5. Guided-TTS with Transribed Speech Data

Guided-TTS leverages the phoneme classifier trained with LibriSpeech for speech synthesis, taking advantage of train-

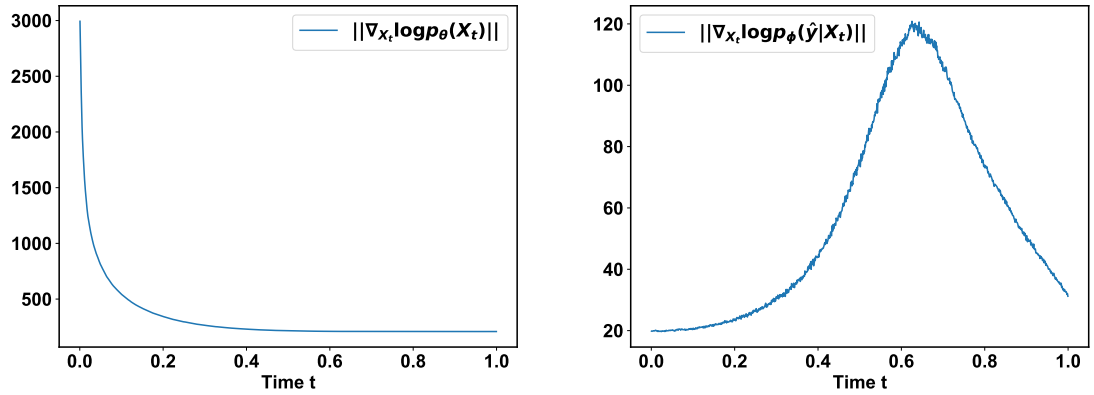

Figure 4: The norm of the unconditional score and the classifier gradient for each timestep t. (Left) The norm of the unconditional score (Right) The norm of the classifier gradient.

ing unconditional DDPM and phoneme classifier separately. If transcripts corresponding to the target speaker's speech exist, we can train the phoneme classifier and duration predictor using the target speaker's dataset instead of using LibriSpeech. We refer to this model as Guided-TTS-T. Since Guided-TTS-T uses the same dataset when training the unconditional DDPM and phoneme classifier, there is no need for the phoneme classifier to generalize to unseen speakers, which makes speaker encoder unnecessary for Guided-TTS-T. For comparison, we train all modules in Guided-TTS-T using LJSpeech.

The unconditional DDPM of Guided-TTS-T is trained using the untranscribed speech of LJSpeech in the same way as the unconditional DDPM of Guided-TTS. The phoneme classifier and duration predictor of Guided-TTS-T use the same structure and hyperparameters used in Guided-TTS. We train the phoneme classifier for 1000 epochs and the duration predictor for 60 epochs. Similar to Guided-TTS, we use the checkpoint that scores best on respective metrics (validation accuracy for phoneme classifier, and validation loss for duration predictor) for evaluation.

The performance and CER of Guided-TTS-T are shown in Table 5. Guided-TTS-T obtains similar performance to Glow-TTS and Grad-TTS even when using the same amount of training data. As demonstrated from this result, Guided-TTS-T is a new approach to construct high-quality TTS in a situation where the target speaker's transcribed data is given.

A.6. Inpainting

We perform the inpainting task to show how well the unconditional DDPM learns the dependencies in mel-spectrogram. The pre-trained unconditional DDPM fills out the masked part of the mel-spectrogram. We use samples from three speakers; one female speaker (ID: 92), one male speaker (ID: 6097) from Hi-Fi TTS, and a female speaker from

Table 5: Mean Opinion Score (MOS) with 95% confidence intervals of TTS models for LJSpeech. "GT MEL" represents the HiFi-GAN result of ground truth mel-spectrogram.

Method	5-scale MOS	CER(%)
GT	4.45 ± 0.05	0.64
GT MEL	4.24 ± 0.07	0.77
GLOW-TTS	4.14 ± 0.08	0.66
GRAD-TTS	4.25 ± 0.07	1.09
GUIDED-TTS-T	4.23 ± 0.08	1.21

LJSpeech. Two cross-shaped masks (LJSpeech, Hi-Fi TTS male) and one binarized MNIST (LeCun & Cortes, 2010) mask (Hi-Fi TTS female) are used for masking. We set 1000 as the number of reverse steps N and $\tau = 1.5$ for inpainting. The method of inpainting is the same as Song et al. (2021b), and the algorithm is as follows:

Algorithm 2 Inpainting Mel-spectrogram

Binary Mask: M, Original mel-spectrogram: $\hat{X}_0$
θ: parameter of unconditional DDPM
$X_1 \sim \mathcal{N}(0, \tau^{-1} I)$
for $i = N$ **to** 1 **do**
  $t \leftarrow \frac{i}{N}$
  $\rho(\hat{X}_0, t) \leftarrow \mathrm{e}^{-\frac{1}{2}\int_0^t \beta_s ds} \hat{X}_0$
  $\lambda(t) \leftarrow I - \mathrm{e}^{-\int_0^t \beta_s ds}$
  $\hat{X}_t \sim \mathcal{N}(\rho(\hat{X}_0, t), \lambda(t))$
  $X_t \leftarrow X_t \odot M + \hat{X}_t \odot (1 - M)$
  $z_t \sim \mathcal{N}(0, \tau^{-1} I)$
  $X_{t-\frac{1}{N}} \leftarrow X_t + \frac{\beta_t}{N}(\frac{1}{2} X_t + \nabla_{X_t} \log p_\theta(X_t)) + \sqrt{\frac{\beta_t}{N}} z_t$
end for
return $X_0 \odot M + \hat{X}_0 \odot (1 - M)$

The inpainting results are shown in Fig. 5, where (a) is the original mel-spectrogram, (b) is the masked mel-

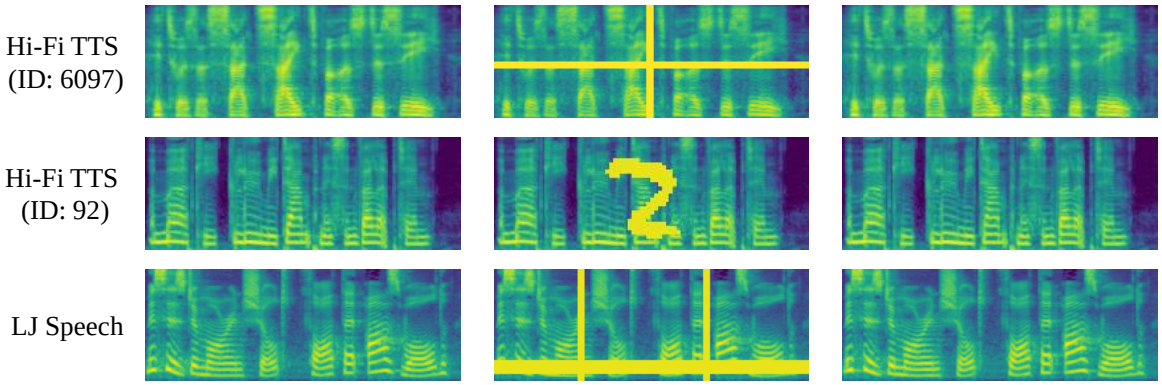

Figure 5: Mel-spectrogram inpainting results of unconditional DDPM trained on LJSpeech, and two speakers (Speaker ID: 92, 6097) from Hi-Fi TTS.

spectrogram, and (c) is the result of inpainting on the masked part. As shown in Fig. 5, we show that the unconditional DDPM of Guided-TTS learns the adjacent frequency and temporal dependencies of the mel-spectrogram. Samples of inpainting results are provided on the demo page.