GUIDED-TTS: TEXT-TO-SPEECH WITH UNTRANSCRIBED SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Most neural text-to-speech (TTS) models require ⟨speech, transcript⟩ paired data from the desired speaker for high-quality speech synthesis, which limits the usage of large amounts of untranscribed data for training. In this work, we present Guided-TTS, a high-quality TTS model that learns to generate speech from untranscribed speech data. Guided-TTS combines an unconditional diffusion probabilistic model with a separately trained phoneme classifier for text-to-speech. By modeling the unconditional distribution for speech, our model can utilize the untranscribed data for training. For text-to-speech synthesis, we guide the generative process of the unconditional DDPM via phoneme classification to produce mel-spectrograms from the conditional distribution given transcript. We show that Guided-TTS achieves comparable performance with the existing methods without any transcript for LJSpeech. Our results further show that a single speaker-dependent phoneme classifier trained on multispeaker large-scale data can guide unconditional DDPMs for various speakers to perform TTS.

1 INTRODUCTION

Neural text-to-speech (TTS) models have achieved to generate high quality human-like speech given text (van den Oord et al. (2016); Shen et al. (2018)). In general, these TTS models are conditional generative models that encode text into the hidden representation and generate speech from the encoded representation. Early TTS models are autoregressive generative models which generate high-quality speech, but suffer from 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 are capable of generating high-quality speech with a faster inference speed (Ren et al. (2019); Ren et al. (2021); Kim et al. (2020); Popov et al. (2021a)). Recently, high-quality end-to-end tts models have been proposed that generate raw waveform from text at once (Kim et al. (2021); Weiss et al. (2021); Chen et al. (2021b)).

Despite the high quality and fast speech synthesis, most TTS models can only be trained with the transcribed data of the desired speaker. While a lot of long-form untranscribed data, such as audiobooks or podcasts, is available in website, it is challenging to make use of these unpaired speech data for training existing TTS models. To exploit these untranscribed data, long-form untranscribed speech data has to be segmented into sentence-level, and each segmented speech should be transcribed accurately. The reason why it is difficult to use untranscribed data is that the existing TTS models directly model the conditional distribution given text.

In this work, we propose Guided-TTS, an unconditional diffusion-based generative model trained on untranscribed data that leverages phoneme classifier for text-to-speech synthesis. To utilize a large amount of untranscribed data for TTS, we train an unconditional diffusion probabilistic model which learns to generate mel-spectrograms without any context. For unconditional speech modeling, training data does not have to be aligned with text sequence. Thus we simply use random chunks of untranscribed speech as the training data for our DDPM, which allows us to use long-form untranscribed data for training without extra effort.

In order to guide the unconditional DDPM for TTS, we train a framewise phoneme classifier on transcribed data and use the gradients of the classifier during sampling. Although our generative model is trained without any transcript, Guided-TTS effectively generates mel-spectrograms given the transcript by guiding the generative process of unconditional DDPM with the phoneme classifier.

We demonstrate that the proposed method, TTS by guiding the unconditional DDPM, matches the performance of the existing conditional TTS models on LJSpeech dataset. We further show that by training phoneme classifier on available multispeaker transcribed dataset, Guided-TTS also obtains comparable performance without seeing any transcript of LJSpeech, which shows the possibility to construct high quality text-to-speech model wihtout transcript. We encourage the readers to listen to samples from 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 (Ho et al. (2020)), which is recently proposed as a kind of probabilistic generative model, has been applied to various domains such as image (Dhariwal & Nichol (2021)), and audio (Chen et al. (2021a); Popov et al. (2021a)). DDPM first defines a forward process that gradually corrupts the data X_0 to a random noise X_T across T timesteps. To generate data from random noise, the model learns the reverse process that follows the reverse trajectory of the predefined forward process.

Recently, there have been approaches to formulate the trajectory between data and noise as a continuous stochastic differential equation (SDE) instead of a discrete-time Markov process (Song et al. (2021b)). Grad-TTS (Popov et al. (2021a)) introduce SDE formulation to TTS, and we follow the formulation of it. Grad-TTS defines a SDE that can generate data X_0 from random noise X_T sampled from $\mathcal{N}(\mu, \Sigma)$. According to the formulation of Grad-TTS, the forward process that corrupts data X_0 into the 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) showed that reverse process, which represents the trajectory from noise X_T to X_0, can be formulated in SDE, which is defined as below:

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

Given the score, the gradient of log density w.r.t 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). In order to generate data, DDPM learns to estimate the score with the neural network s_θ parameterized by θ.

To estimate the score obtained from the forward process Eq. (1), 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 $\lambda(t) = I - \mathrm{e}^{-\int_0^t \beta_s ds}$. Then, the score can be derived from Eq. (3); $\nabla_{X_t}\log p_t(X_t|X_0) = -\lambda(t)^{-1}\epsilon_t$ given X_0 (Popov et al. (2021a)). 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 L2 loss as in previous works (Ho et al. (2020), Song et al. (2021b)).

Using the 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 to solve SDE, $t \in \{\frac{1}{N}, \frac{2}{N}, ..., 1\}$ and z_t is a standard Gaussian noise.

[1] Demo : https://anonymousauthors2022.github.io/

2.2 CLASSIFIER GUIDANCE

DDPM can be guided to generate samples with desired condition without fine-tuning by introducing the classifier. Song et al. (2021b) used the unconditional DDPM to generate class-conditional images using separately trained image classifier. Not only unconditional DDPM but also conditional DDPM can be guided using the classifier, which makes the usability of DDPM wider and contributes to achieving state-of-the-art performance in the class conditional image generation (Dhariwal & Nichol (2021)).

For conditional generation, the classifier $p_t(y|X_t)$ is trained to classify the noisy data X_t as the condition y. If Eq. (5) is modified, discretized SDE for the conditional generation can be obtained.

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

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

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

3 GUIDED-TTS

In this section, we present Guided-TTS. Our goal is to build a high-quality text-to-speech model that leverages a large amount of untranscribed data for training. While other TTS models directly learn to generate speech from text, Guided-TTS models the unconditional distribution of speech to utilize speech-only data, and guides the unconditional model to generate speech with a given text. To the best of our knowledge, Guided-TTS is the first TTS model that generates speech with unconditional generative model.

Both the unconditional speech modeling and controllable generation with the unconditional generative model are well known to be challenging. To tackle these challenges, we adopt a diffusion-based generative model for unconditional speech generation, which has advantages of modeling complex distribution and easy controllability. Additionally, we introduce a phoneme classifier to guide the unconditional DDPM for TTS.

As the generative model and the classifier are trained separately, we can use different training data for each model. Therefore, even if only the untranscribed speech data is available for TTS, the phoneme classifier can leverage the available transcribed data. By guiding with the phoneme classifier trained on a large-scale multispeaker data, Guided-TTS allows us to achieve the text-to-speech model with the untranscribed speech data.

Guided-TTS consists of 4 modules, unconditional DDPM, phoneme classifier, duration predictor and speaker encoder, as shown in the Fig. 1. Unconditional DDPM is a module that learns to generate mel-spectrogram unconditionally, and the remaining three modules are for TTS synthesis by guidance. We will explain the unconditional DDPM in Section 3.1, and the way to guide 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 assume that the training data for the diffusion-based model has tens of hours speech from a single speaker S. As our generative model learns only with speech, training samples do not need to be aligned with text. Thus, we simply use random chunks of untranscribed speech as training data so that we can reduce the burden of not only speech transcription but also segmentation even in the form of long-form speech data.

Given a mel-spectrogram $X = X_0$, we define a 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 estimates a score with the neural network $s_\theta(X_t, t)$ parameterized by θ. The training objective of our unconditional model is in Eq. (4).

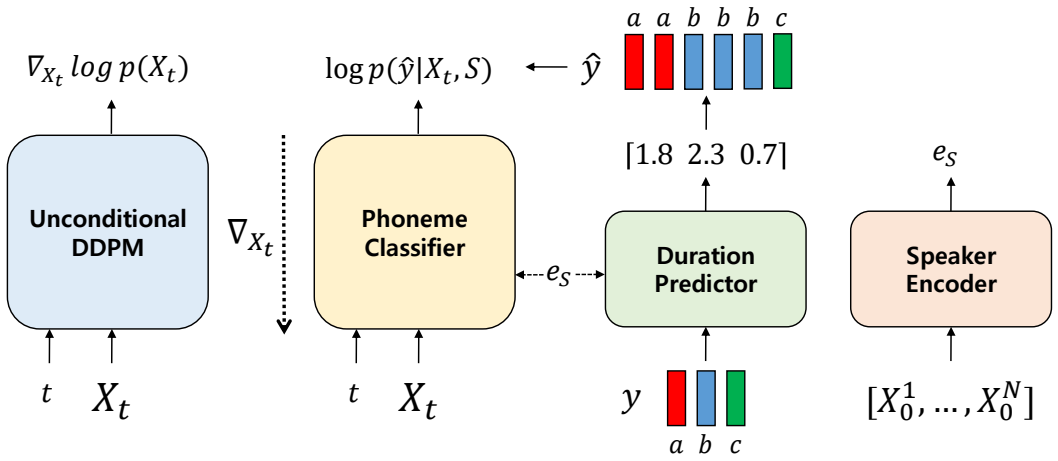

Figure 1: The overall architecture of Guided-TTS. The unconditional DDPM learns to generate speech X_0 without transcript. The other modules, the phoneme classifier, duration predictor, and speaker encoder are for guiding the unconditional DDPM to generate conditional samples given y. The speaker embedding e_S is only used for training speaker-dependent phoneme classifier.

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

In general, it is difficult to evaluate unconditional speech generative models. Since the goal of this work is not unconditional spoken language modeling but text-to-speech synthesis, we focus on demonstrating the performance of the unconditional DDPM in a few interpretable tasks: inpainting and TTS. The task of mel-spectrogram inpainting is to fill a masked part in a mel-spectrogram, which we can indirectly evaluate how well the unconditional DDPM models the dependencies of speech. We can perform the mel-spectrogram inpainting task by iteratively refining the masked part of the mel-spectrogram using the unmasked part (Song et al. (2021b)), which we describe in the Appendix A.1.

3.2 Text-to-Speech via Classifier Guidance

For TTS synthesis, we use a framewise phoneme classifier to guide the unconditional DDPM. 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 $\hat{y}$. Then, we sample a random noise X_T of the same length as $\hat{y}$ from the standard normal distribution and we can generate conditional samples by replacing the unconditional score in Eq. (5) with the conditional score. The conditional score on the left side can be estimated 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 with the phoneme classifier, as in Eq. (8). That is, we achieve to 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.

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

If the transcripts are available for the speech data of the target speaker S, we train both the unconditional DDPM and the phoneme classifier on the same data from the speaker S for TTS synthesis. Otherwise, to guide unconditional DDPM trained on untranscribed data for the speaker S, we train the phoneme classifier on multi-speaker transcribed data to generalize to the unseen speaker S. We further provide the speaker embedding extracted from the pretrained speaker verification network to

the phoneme classifier to better guide the unconditional DDPM trained on the unseen speaker S. In order to predict the duration for the speaker S, the speaker embedding is also provided as a condition to the duration predictor. We describe each module required for guiding below.

Phoneme Classifier The phoneme classifier is a network trained on transcribed data that recognizes the phoneme corresponding to each frame of the input mel-spectrogram. For training framewise phoneme classifier, we align transcript and speech using a forced alignment tool, Montreal Forced Aligner (MFA) (McAuliffe et al. (2017)), and extract the frame-level phoneme label $\hat{y}$. To better guide the generative process of diffusion-based generative model, the phoneme classifier is trained to classify the corrupted mel-spectrogram X_t sampled from (Eq 2) as the frame-level phoneme label $\hat{y}$. The training objective of phoneme classifier is to maximize the expectation of cross entropy between the phoneme label $\hat{y}$ and the output probability with respect to $t \in [0, 1]$.

We use 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. (2021a), is used as a global condition in WaveNet to provide information about the noise level of the corrupted input X_t at timestep t. For speaker-dependent classification, we additionally use the speaker embedding e_S from the speaker encoder as the global condition.

Duration Predictor 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 that the phoneme classifier is trained on. The duration predictor is trained to minimize L2 loss between the duration label and the estimated duration in log-domain, and we round up the estimated duration during inference. The architecture of the duration predictor is same as that of Glow-TTS (Kim et al. (2020)) with the text encoder.

Speaker Encoder 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 speaker encoder to condition speaker-dependent modules. For training speaker-dependent modules, we use speaker embedding e_S extracted from the clean mel-spectrogram X_0 for each training data. For guiding, we average and normalize the speaker embeddings of untranscribed speech for the desired speaker S to extract e_S.

3.2.1 NORM-BASED GUIDANCE

Algorithm 1 Norm-based Guidance

$\hat{y}$: framewise phoneme label, s: gradient scale
$X_1 \sim \mathcal{N}(0, I)$
for $i = N$ **to** 1 **do**
  $t \leftarrow \frac{i}{N}$
  $\alpha_t \leftarrow \|\nabla_{X_t} \log p_t(X_t)\| \,/\, \|\nabla_{X_t} \log p_t(\hat{y}|X_t)\|$
  $z_t \sim \mathcal{N}(0, I)$
  $X_{t-\frac{1}{N}} \leftarrow X_t + \frac{\beta_t}{N}(\frac{1}{2}X_t + \nabla_{X_t} \log p_t(X_t) + s \cdot \alpha_t \nabla_{X_t} \log p_t(\hat{y}|X_t))) + \sqrt{\frac{\beta_t}{N}} z_t$
end for
return X_0

When guiding the unconditional DDPM with our phoneme classifier, we found that the norm of the unconditional score suddenly increases near the data. That is, the closer to the data, the phoneme classifier has little effect on the generative process of DDPM. Here, we propose norm-based guidance to better guide the unconditional DDPM to generate 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 disappearing as the score rises steeply. 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.

4 EXPERIMENTS

Datasets Since our proposed method basically separates the training of unconditional DDPM and the phoneme classifier, they can be trained using different datasets. To show the performance of Guided-TTS on transcribed data, we used LJSpeech dataset for training both the unconditional model and the classifier. In addition, we demonstrate the performance of Guided-TTS with untranscribed data by using only the speech data from LJSpeech dataset, Hi-Fi TTS dataset (Bakhturina et al. (2021)), and Blizzard 2013 dataset (King & Karaiskos (2013))). In this case, we used these speech only data for training unconditional DDPMs, and to guide the various unconditional generative models, we trained a speaker-dependent phoneme classifier on LibriTTS dataset (Zen et al. (2019)), a multi-speaker large-scale dataset with approximately 585 hours of speech uttered by 2456 speakers with corresponding text. To extract speaker embedding from each utterance, we trained speaker encoder on VoxCeleb2 dataset (Chung et al. (2018)), a speaker verification dataset that contains more than 1M utterances for 6112 speakers.

LJSpeech is a 24-hour single female speaker dataset consisting of 13,100 audio clips. Dataset are randomly splitted, 12,500 samples for training set, 100 samples for validation set, and 500 samples for test set. Hi-Fi TTS dataset is a multispeaker dataset with 6 females and 4 males and each speaker's data consists of at least 17 hours of speech. We select three of them (two males and one female) and used them to train three unconditional DDPM, respectively. The Blizzard 2013 dataset is an 147 hours segmented and unsegmented audiobook data which is read by a single female speaker. We only used 5 second randomly clipped audio of unsegmented data to show that our model does not require text labeling or text aligning for untranscribed dataset TTS. Only audio file is used to train unconditional DDPM.

Training Details Text is converted into International Phonetic Alphabet (IPA) phoneme sequences using open-source software (Bernard (2021)). To extract the mel-spectrogram, we use the same hyperparameters as the official implementation of Glow-TTS (Kim et al. (2020)).

For all models, Adam optimizer was used, the learning rate was 0.0001, and the betas for optimizer were (0.9, 0.999). For the unconditional model and the Phoneme classifier, $\beta_0 = 0.05, \beta_1 = 20$ were used for beta schedule. For unconditional model, the base model had the same architecture and hyperparameters as DDPM (Ho et al. (2020)) used for 32×32 image modeling, and for ablation, we used the same architecture and hyperparameters for small model as Grad-TTS (Popov et al. (2021a)). Other details and hyperparameters are described in the Appendix A.2.

Evaluation To compare the performance of models, pretrained models and the public implementations of Glow-TTS and Grad-TTS were used.[2] For Grad-TTS and Guided-TTS, we set N, the number of reverse steps to 50. For comparison of TTS performance results on the LJSpeech dataset, HiFi-GAN (Kong et al. (2020)) fine-tuned to the mel-spectrogram generated by Tacotron 2 (Shen et al. (2018)) was used as a vocoder. When evaluating TTS performance for other datasets, we used universal HiFi-GAN for vocoder.

Inpainting was performed to show how well the unconditional DDPM models the mel-spectrogram dependency. Three speakers, one female speaker and one male speaker from HiFi-TTS dataset, and the speaker from LJSpeech dataset were used, and two cross shaped masks and one binarized MNIST (LeCun & Cortes (2010)) mask were used for masking. The number of reverse steps we used for inpainting is 1000.

5 RESULTS

5.1 MODEL COMPARISON

We compared the performance of audio samples by measuring the 5-scale mean opinion score (MOS) using Amazon Mechanical Turk. 50 samples were randomly selected from the LJSpeech

[2]The implementations are as follows:
HiFi-GAN : https://github.com/jik876/hifi-gan
Glow-TTS : https://github.com/jaywalnut310/glow-tts
Grad-TTS : https://github.com/huawei-noah/Speech-Backbones/tree/main/Grad-TTS

test set and the results are shown in Table 1.[3] The quality of Guided-TTS with transcribed LJSpeech dataset (4.18 ± 0.07) shows that our model achieves comparative TTS performance to Glow-TTS and Grad-TTS which directly model conditional distribution.

Unlike Grad-TTS, which models conditional distribution, our unconditional DDPM requires a large model size as long-term dependency of mel-spectrogram have to be modeled without text. The MOS result of small model, whose architecture and hyperparameters are the same as GradTTS, is 4.01 ± 0.07. The lower score with small unconditional DDPM demonstrates the importance of modeling capability of unconditional DDPM.

For transcribed data, Guided-TTS matches the performance of existing TTS with conditional speech modeling. In addition, we show that we can generate high quality samples using untranscribed speech of speaker, which is comparative to the samples from other TTS models which is trained using speech, text pair of desired speaker (4.18 ± 0.07). This demonstates that Guided-TTS enables high quality TTS using untranscribed speech. Samples of all models used for comparison are available on the demo page.

Table 1: Mean Opinion Score (MOS) of TTS models on LJSpeech with 95% confidence intervals.

Method	5-scale MOS
Ground Truth	4.46 ± 0.05
Ground Truth (Mel + HiFi-GAN (Kong et al. (2020)))	4.25 ± 0.07
Glow-TTS (Kim et al. (2020))	4.10 ± 0.10
Grad-TTS (Popov et al. (2021a))	4.25 ± 0.09
Guided-TTS (LJ base, w/ LJ Phoneme Classifier)	4.18 ± 0.07
Guided-TTS (LJ small, w/ LJ Phoneme Classifier)	4.01 ± 0.07
Guided-TTS (LJ base, w/ LibriTTS Phoneme Classifier)	4.18 ± 0.07

Table 2: Mean Opinion Score (MOS) of Guided-TTS with various untranscribed datasets with 95% confidence intervals.

Method	5-scale MOS	Total duration (hrs)	Gender
Guided-TTS (LJSpeech)	4.18 ± 0.07	24	Female
Guided-TTS (HiFi-TTS ID: 92)	3.97 ± 0.08	27.3	Female
Guided-TTS (HiFi-TTS ID: 6097)	3.82 ± 0.08	30.1	Male
Guided-TTS (HiFi-TTS ID: 9017)	3.72 ± 0.08	58.0	Male
Guided-TTS (Blizzard 2013)	3.85 ± 0.08	149.4	Female

5.2 Generalization to Diverse Datasets

We show that our model can synthesize high quality speech using untranscribed speech. Since our model trains the unconditional model and the classifier separately, TTS for various untranscribed datasets is possible with a single classifier trained with a large scale multispeaker dataset. The performance of our multiple TTS model using untranscribed speech are presented in Table 2. Our model generates high quality sample by guiding unconditional DDPM even on text-free data. In particular, high TTS performance was achieved even for randomly cropped unsegmented Blizzard 2013 dataset. Our model is capable of high quality TTS for easily available untranscribed speech such as audiobooks without transcription and segmentation.

Our method enables TTS for diverse untranscribed datasets with various characteristics (*e.g.,* gender, accent, prosody). In addition to TTS performance for multiple datasets in Table 2, there are samples in the demo page for diverse datasets, so we highly encourage the reader to listen to it.

[3]Demo : https://anonymousauthors2022.github.io/

5.3 INPAINTING

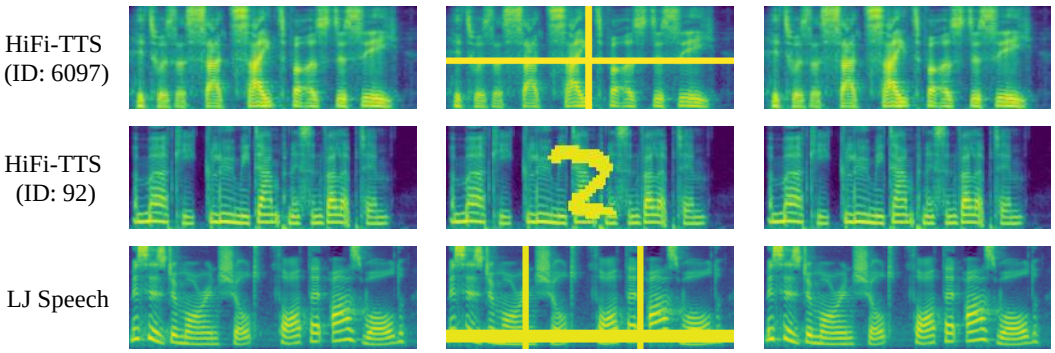

Figure 2: Mel-spectrogram Inpainting Results of Unconditional DDPM trained on LJSpeech, and two speakers (Speaker ID: 92, 6097) from HiFi-TTS.

In a previous session, we showed that our TTS model is comparative with existing works and is capable of high quality TTS using a variety of untranscribed speech. We additionally performed inpainting to demonstrate how well our unconditional DDPM performs mel-spectrogram modeling on its own. The inpainting results for multiple datasets are shown in Fig. (2). Through these mel-spectrogram inpainting results, we show that our unconditional DDPM models the adjacent frequency and temporal dependencies of the mel-spectrogram well. Unconditional and inpainting samples generated by our model are provided on the sample pages.

6 RELATED WORK

Unconditional Speech Generation In general, the unconditional generative model (van den Oord et al. (2016); Vasquez & Lewis (2019)), which uses only raw waveforms for training, is more difficult to learn than the conditional generative model that synthesizes speech using text or mel-spectrograms. Several papers attempt to unconditionally generate raw waveforms (van den Oord et al. (2016); Donahue et al. (2019)), however, they have difficuly in capturing very large receptive fields, and thus generate samples of poor quality. Previous works (van den Oord et al. (2017); Vasquez & Lewis (2019)) that attempt to model the latent code or mel-spectrogram of audio instead of directly modeling raw waveforms require modeling a relatively small receptive field, so the burden of the model is less, ensuring a high quality sample. Almost all unconditional audio (or mel-spectrogram) generation models that exist so far have been used for unconditional audio modeling ,but not for any other purpose. To the best of our knowledge, this is the first application of an unconditional model to a text-to-speech (TTS) model with appropriate guidance to enable the speech synthesis of untranscribed data.

Text-to-Speech Models Most text-to-speech (TTS) models are composed of two parts: a model taht generates intermediate features (*e.g.,* mel-spectrogram) from text (Shen et al. (2018)) and the vocoder, which synthesizes raw waveforms from intermediate features (van den Oord et al. (2016)). The autoregressive generative 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. Since this autoregressive model sequential generates a sample, the inference speed is slow. Parallel TTS models using various generative models have been proposed to solve the problem. Flow-based generative model (Kingma & Dhariwal (2018)) and feed-forward model have been proposed for text-to-mel-spectrogram models (Ren et al. (2019); Ren et al. (2021); Kim et al. (2020); Shih et al. (2021)) and vocoders (Oord et al. (2018); Prenger et al. (2019); Kim et al. (2019)) which enable fast sampling. Similarly, Variational Autoencoder (Kingma & Welling (2014)) based models (Lee et al. (2020); Liu et al. (2021)), Denoising Diffusion Probabilistic Models (Ho et al. (2020)) based models (Chen et al. (2021a); Kong et al. (2021); Popov et al. (2021a); Jeong et al. (2021)), GAN (Goodfellow et al. (2014)) based mod-

els (Kumar et al. (2019); Bińkowski et al. (2019); Kong et al. (2020)) are TTS models with parallel sampling schemes.

Recently, unlike the existing two-stage-based models, end-to-end TTS models have been proposed, such as FastSpeech2s (Ren et al. (2021)), EATS (Donahue et al. (2021)), Wave-Tacotron (Weiss et al. (2021)), VITS (Kim et al. (2021)), and WaveGrad2 (Chen et al. (2021b)). Most two-stage TTS models and end-to-end TTS models perform conditional generation tasks with models trained using text and speech pairs. We proposed a new TTS model that learns the speaker's speech using untranscribed data and guides it through a classifier trained with other datasets that have text and speech pairs. By guiding an unconditional model with an independently pretrained phoneme classifier, a high-performance TTS model of the desired speaker can be trained with much less effort.

Diffusion-based Generative Models DDPM (Ho et al. (2020)) has undergone several theoretical developments (Song et al. (2021b)) and has been used in many domains to produce high quality samples (Ho et al. (2020); Dhariwal & Nichol (2021); Chen et al. (2021a); Popov et al. (2021a); Luo & Hu (2021)). Many theoretical and practical breakthrougsh such as better noise schedules (Nichol & Dhariwal (2021); Lam et al. (2021)), faster sampling (Chen et al. (2021a); Song et al. (2021a); Watson et al. (2021); Kong & Ping (2021); Jolicoeur-Martineau et al. (2021)) have been proposed, and a continuous version of DDPM, an SDE-based model (Song et al. (2021b) ; Popov et al. (2021a)) is also presented. In the audio domain, DDPM has shown strong performance in many tasks such as Vocoder (Chen et al. (2021a); Kong et al. (2021)), and voice conversion (Popov et al. (2021b)). In particular, GradTTS (Popov et al. (2021a)) and Diff-TTS (Jeong et al. (2021)) generated mel-spectrograms using DDPM when text is provided.

Unlike other generative models, DDPM uses a pretrained unconditional model for various tasks such as imputation (Song et al. (2021b)), manipulation (Choi et al. (2021); Meng et al. (2021)), and controllable generation (Song et al. (2021b)). In particular, the controllable generation allows the DDPM to achieve state-of-the-art performance in image conditional generation by guiding an unconditional sample to the desired condition using a gradient from the separately trained classifier (Dhariwal & Nichol (2021)). We used the DDPM architecture to perform TTS by guiding the unconditional model trained without transcript, unlike the TTS model that could not be trained on existing untranscribed data. By using DDPM for untranscribed data TTS, powerful controllable generation capability can be leveraged.

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, the proposed model can leverage a large amount of untranscribed data for training. Thanks to the properties of diffusion based generative models, our unconditional generative model can generate speech given transcript by introducing the separately trained phoneme classifier. 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 to the previous TTS models on LJSpeech dataset. Moreover, Guided-TTS without transcript generates samples comparable to existing models using transcripts by training its phoneme classifier on a large-scale multispeaker dataset. We also showed that the single well-performed phoneme classifier can guide various unconditional DDPMs to generate high quality sample. We believe that Guided-TTS can reduce the burden of constructing training dataset for high quality TTS.

References

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

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*, 2021.

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

Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High Fidelity Speech Synthesis with Adversarial Networks. In *International Conference on Learning Representations*, 2019.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*, 2021a.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis. In *Proc. Interspeech 2021*, pp. 3765–3769, 2021b. doi: 10.21437/Interspeech.2021-1897.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

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

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

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

Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, and Karen Simonyan. End-to-end Adversarial Text-to-Speech. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rsf1z-JSj87.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 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.

Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. In *Proc. Interspeech 2021*, pp. 3605–3609, 2021. doi: 10.21437/Interspeech.2021-469.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.

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

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

Sungwon Kim, Sang-Gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. Flowavenet: A generative flow for raw audio. In *International Conference on Machine Learning*, pp. 3370–3378, 2019.

Simon J. King and Vasilis Karaiskos. The blizzard challenge 2013. In *In Blizzard Challenge Workshop*, 2013.

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

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

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

Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021. URL https://openreview.net/forum?id=agj4cdOfrAP.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*, 2021.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems 32*, pp. 14910–14921, 2019.

Max WY Lam, Jun Wang, Rongjie Huang, Dan Su, and Dong Yu. Bilateral denoising diffusion models. *arXiv preprint arXiv:2108.11514*, 2021.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

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

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6706–6713, 2019.

Peng Liu, Yuewen Cao, Songxiang Liu, Na Hu, Guangzhi Li, Chao Weng, and Dan Su. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. *arXiv preprint arXiv:2102.06431*, 2021.

Shitong Luo and Wei Hu. 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.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*, 2017.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (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.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *International Conference on Learning Representations*, 2018.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 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, 2021a.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821*, 2021b.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 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.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, Robust and Controllable Text to Speech. volume 32, pp. 3171–3180, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 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.

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

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, 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.

Kevin J Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. Radtts: 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.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

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

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

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 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.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pp. 4006–4010, 2017. doi: 10.21437/Interspeech.2017-1452.

Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021.

Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P Kingma. 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.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

A APPENDIX

A.1 INPAINTING ALGORITHM

In this section, we would like to describe the algorithm of inpainting. The method of performing 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: $\hat{X}_0$
$X_1 \sim \mathcal{N}(0, I)$
for $i = N$ **to** 1 **do**
  $t \leftarrow \frac{i}{N}$
  $\rho(\hat{X}_0, t) \leftarrow \mathrm{e}^{-\frac{1}{2}\int_0^t \beta_s ds}\hat{X}_0$
  $\lambda(t) \leftarrow I - \mathrm{e}^{-\int_0^t \beta_s ds}$
  $\hat{X}_t \sim \mathcal{N}(\rho(\hat{X}_0, t), \lambda(t))$
  $X_t \leftarrow X_t \odot M + \hat{X}_t \odot (1 - M)$
  $z_t \sim \mathcal{N}(0, I)$
  $X_{t-\frac{1}{N}} \leftarrow X_t + \frac{\beta_t}{N}(\frac{1}{2}X_t + \nabla_{X_t} \log p_t(X_t)) + \sqrt{\frac{\beta_t}{N}} z_t$
end for
return $X_0 \odot M + \hat{X}_0 \odot (1 - M)$

A.2 TRAINING DETAILS AND HYPERPARAMTERS

In this section, we cover the training details and hyperparameters which are not mentioned in main text. Among the four models constituting Guided-TTS, unconditional models were trained with batch size 16 for all datasets. The Phoneme classifier uses the WaveNet structure and is trained on the LJSpeech and LibriTTS datasets. The classifier trained in LJSpeech dataset used 256 residual channels, 6 residual blocks stacks of 3 dilated convolution layers, and was trained for 1000 epochs. For LibriTTS dataset, we used 512 residual channels, 3 residual blocks stacks of 6 dilated convolution layers for classifier which was trained for 100 epochs. Duration predictor was trained with batch size 64. For sampling, we use the checkpoint of the best epoch with the best metric for phoneme classifier (validation accuarcy) and duration predictor (validation loss).