Mandarin Singing Voice Synthesis with Denoising Diffusion Probabilistic Wasserstein GAN

Yin-Ping Cho*, Yu Tsao†, Hsin-Min Wang‡, and Yi-Wen Liu*
* National Tsing Hua University, Hsinchu, Taiwan
Tel/Fax: +886-3-5162205
† Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: yinping.cho@outlook.com, yu.tsao@citi.sinica.edu.tw, whm@iis.sinica.edu.tw, ywliu@ee.nthu.edu.tw

***Abstract*—Singing voice synthesis (SVS) is the computer production of a human-like singing voice from given musical scores. To accomplish end-to-end SVS effectively and efficiently, this work adopts the acoustic model-neural vocoder architecture established for high-quality speech and singing voice synthesis. Specifically, this work aims to pursue a higher level of expressiveness in synthesized voices by combining the diffusion denoising probabilistic model (DDPM) and *Wasserstein* generative adversarial network (WGAN) to construct the backbone of the acoustic model. On top of the proposed acoustic model, a HiFiGAN neural vocoder is adopted with integrated fine-tuning to ensure optimal synthesis quality for the resulting end-to-end SVS system. This end-to-end system was evaluated with the multi-singer Mpop600 Mandarin singing voice dataset. In the experiments, the proposed system exhibits improvements over previous landmark counterparts in terms of musical expressiveness and high-frequency acoustic details. Moreover, the adversarial acoustic model converged stably without the need of enforcing reconstruction objectives, which shows the convergence stability of the proposed DDPM and WGAN combined architecture over alternative GAN-based SVS systems.**[1]

I. Introduction

Singing voice synthesis (SVS) aims to generate singing voices as natural and expressive as those of human singers. The attention from both academia and commercial corporations has pushed the boundaries of SVS with neural networks in recent years.

Although there are neural network-driven SVS systems designed to generate waveforms directly from the musical scores [1], [2], the most prominent system design is to split the generation pipeline into two stages. First, an *acoustic model* frontend consumes the musical score to estimate the intermediate acoustic features for the singing voice; then, a backend synthesizes the final audio waveform from those acoustic features. While the intermediate acoustic features can be spectrograms [3] or the *WORLD* vocoder [4] parameters [5], [6], Mel-spectrograms are preferred by most neural SVS systems pursuing the highest audio quality as in [7]–[9]. These Mel-spectrogram intermediates are transformed to audio waveforms with neural vocoder backends such as Parallel WaveGAN [10], HiFiGAN [11], or the novel singing-specific SawSing [12] neural vocoder.

In this setup, the acoustic model decides how musically natural and expressive the synthesized singing voice will be. Numerous deep learning techniques have been deployed as the backbone module in constructing acoustic models. Some of the noticeable techniques in the landmark systems include long-short term memory machine (LSTM) [13], [14], WaveNet [1], [6], and Transformer [5], [7]. These systems commonly employ reconstructive L1 or L2 loss in training the acoustic model to estimate the acoustic features. Nevertheless, generative models trained on simple reconstructive targets often suffer from over-smoothing and producing estimations that approach the mean/median of the target distribution but lack human-like variations.

Therefore, some of the recent advancements in SVS acoustic models are based on generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs). Regarding musical naturalness and expressiveness, GANs have proven to be an enhancement in multiple SVS systems [3], [8], [15], but they are usually less stable in training and may require dataset-specific hyperparameter tuning to converge. On the other hand, DDPM is a generative model that set state-of-the-art performance in recent years across domains such as image generation [16], [17], neural vocoding [18], and speech or singing voice synthesis [9], [19]. With the two techniques being high-performance, their confluence came as a natural development. The combination of GAN and DDPM has shown excellent performances in image generation [20] with a subsequent attempt in speech synthesis [21].

To explore the practicality of employing such a combined architecture of DDPM and GAN, this work trains the DDPM acoustic model adversarially with a *Wasserstein* GAN (WGAN) algorithm. Among different kinds of GANs, WGAN is the one that promises to approximate the real data distribution even when this probability density may not be approachable by other GAN metrics [22], [23]. This property makes it an ideal GAN algorithm for SVS where the complicated singing voice distribution is usually supported by a relatively small dataset due to the difficulties of data

[1]Evaluation audio samples can be found at: https://yinping-cho.github.io/diffwgansvs.github.io/

978-616-590-477-3 ©2022 APSIPA

collection. Additionally, this work proposes a *Musical-Score-Conditioned* (MSC)-discriminator that incorporates information of the sung contents to further prevent mode collapses that are often encountered in GAN-based synthesis models.

II. Denoising Diffusion Wasserstein GAN

The following section introduces DDPM with *Wasserstein GAN* (WGAN) [22]. Section II-A states the denoising diffusion probabilistic model formulation, Section II-B explores why and how a Wasserstein GAN is applied to DDPM, and Section II-C presents the denoising diffusion WGAN we formulate for SVS in this work.

A. Diffusion Denoising Probabilistic Model

The working of DDPM is characterized by a Gaussian Markov chain with a *noising process* and a *denoising process*, which are visualized in Fig. 1.

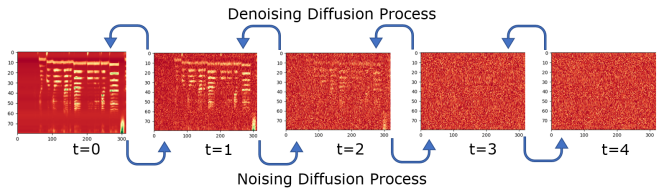


Fig. 1: Visualization of the *noising diffusion process* and the *denoising diffusion process* with a Mel-spectrogram.

The *noising process* transforms a true data sample $\mathbf{x}_0$, in this case a clean Mel-spectrogram, to a unit normal noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ of the same dimensions, formulated as:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &= \prod_{t\geq 1} q(\mathbf{x}_t|\mathbf{x}_{t-1}), t = 1, 2, 3, ..., T; \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \end{aligned} \tag{1}$$

where the variance schedule $\{\beta_t\}$ is pre-defined according to an exponential scheme same as in [20]. Conversely, $q(\mathbf{x}_0)$ denotes the data generation of a true Mel-spectrogram, and $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is thus the true denoising diffusion transition.

The *denoising process* gradually synthesizes an estimated Mel-spectrogram $\hat{\mathbf{x}}_0$ from a random initial noise $\mathbf{x}_T$ with model parameter θ in the form of a *denoising* process:

$$\begin{aligned} p_\theta(\hat{\mathbf{x}}_{0:T}) &= p(\mathbf{x}_T) \prod_{t\geq 1} p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t), t = T, T-1, ...2, 1; \\ p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\hat{\mathbf{x}}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}), \end{aligned} \tag{2}$$

where σ_t^2 is fixed according to each t, and $\mu_\theta(\mathbf{x}_t, t)$ is estimated with conditions $\mathbf{x}_t$, t, and model parameter θ.

Here, same as in [20], we choose to parameterize the denoising transition by first estimating a clean Mel-spectrogram $\hat{\mathbf{x}}_{0,t}$ directly from noised data $\mathbf{x}_t$ and use it to sample the estimated denoised $\hat{\mathbf{x}}_{t-1}$:

$$p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t) := q(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0,t} = f_\theta(\mathbf{x}_t, t)), \tag{3}$$

where the denoising distribution $f_\theta(\mathbf{x}_t, t)$ is estimated by the acoustic model generation $f_\theta(\mathbf{x}_t, t) = G_\theta(\mathbf{x}_t, t, \mathbf{ms}, id)$, with $\mathbf{ms}$ denoting the *musical score* with lyrics, *id* being the *singer identity*, and $G_\theta(\cdot)$ as the neural network generator with learnable parameter θ.

Since the diffusion sampling is simply a Gaussian sampling, $\hat{\mathbf{x}}_t$ can be computed in one step as:

$$\hat{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

and $q(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_0)$ is efficiently obtained with

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}((\mathbf{x}_0, \mathbf{x}_t), \tilde{\beta}_t\mathbf{I})), \\ \tilde{\mu}(\mathbf{x}_0, \mathbf{x}_t) &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \end{aligned} \tag{5}$$

where scalar constants $\alpha_t := 1-\beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, and $\tilde{\beta} := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ can be pre-computed once the variance schedule $\{\beta_t\}$ is defined as a hyperparameter of the process [16], [20].

B. Diffusion Wasserstein GAN

To achieve good synthesis quality with the *denoising process* introduced in Subsection II-A, DDPMs usually set the number of time steps T to a large number (hundreds to thousands) and keep β_t very small. This way, both the *noising* $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the *denoising* $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ diffusion transitions of the true data are sufficiently Gaussian, making it easy for the model to approximate the true denoising transition: $p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

However, setting a large T and small $\{\beta_t\}$ makes the synthesis process much more time-consuming both in the training and the inference phases. Consequently, we adopt a more efficient Variance Preserving (VP) SDE scheme proposed in [24], computing $\{\beta_t\}$ according to

$$\beta_t = 1 - \exp\{\beta_{\min}(\frac{1}{T}) - 0.5(\beta_{\max} - \beta_{\min})\frac{2t-1}{T^2}\}, \tag{6}$$

where we set $\beta_{\max} = 20$, $\beta_{\min} = 0.1$, and number of time steps $T = 4$, found as the best-performing hyperparameters in [20].

Nevertheless, this few-step, big-jump diffusion scheme poses a much greater challenge to the acoustic model. Since the distribution of Mel-spectrograms is apparently not Gaussian and empirically sparse, the diffusion transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is a complicated conditional distribution that is hard to estimate.

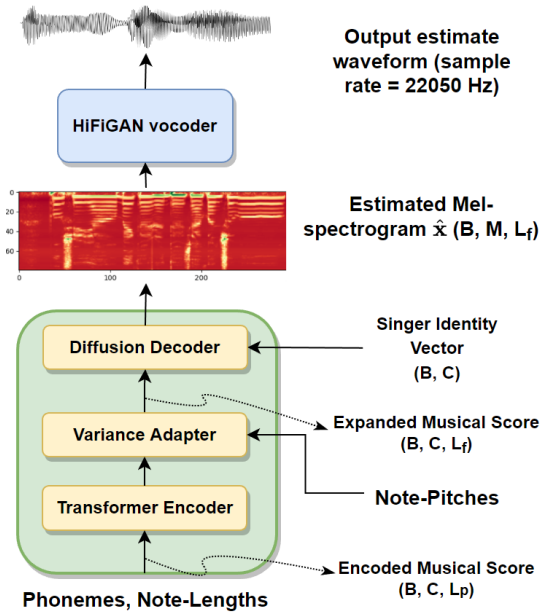
Thus, the objective for the diffusion denoising generator is formulated as

$$\min_\theta \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)}[D_{\text{adv}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t)||p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t))], \tag{7}$$

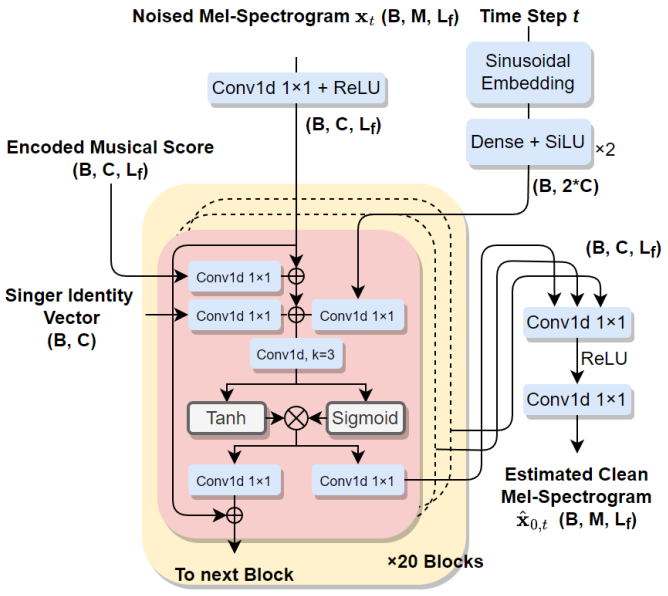
which aims to match the true transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and the model's synthesis transition $p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t)$ by minimizing an *adversarial objective* D_{adv} that estimates the *Wasserstein Distance* with adversarial neural networks.

C. Diff-WGAN formulated for SVS

The *adversarial objective* is constructed by demanding the conditional discriminator to distinguish the true prior Mel-spectrogram $\mathbf{x}_{t-1}$ and the estimated $\hat{\mathbf{x}}_{t-1} \sim p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t)$ generated by the acoustic model per the definition in Equation 3.



(a) Acoustic model overall architecture



(b) DDPM decoder module architecture

Fig. 2: Fig 2a: the overall pipeline of the proposed end-to-end SVS system. Fig 2b: illustration of the WaveNet blocks used in the diffusion decoder. (B: batch size, C: number of channels, L_p: length of the phone sequence, L_f: length of the Mel-spectrogram, M: Mel-spectrograms' number of bins.)

For the SVS task, we formulate the generator and discriminator as:

$$\begin{aligned} &Generator: G_\theta(\mathbf{x}_t, t, \mathbf{ms}, id) \\ &Discriminator: D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t, \mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id) \end{aligned} \tag{8}$$

where $_\theta$ and ϕ denote the learnable parameters for the two networks, $\mathbf{ms}$ denotes the musical score encoded by the musical score encoder and variance adaptor modules, id denotes singer identity, and $\mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}} \in \mathbb{R}^{1\times L_p}$ denote the musical score's phone sequence, note-length sequence, and note-pitch sequence.

Aside from the real/fake $\mathbf{x}_{t-1}$/$\hat{\mathbf{x}}_{t-1}$, the discriminator is provided with a real prior $\mathbf{x}_t$ and the corresponding time step t as completion of conditions for a noising diffusion transition. In addition, the musical score's information and the singer id id are encoded and fed as auxiliary information since they were found to improve training stability and enhance lingual features in related studies [21], [25].

To enforce the *Lipschitz* continuity required by WGAN [22], we applied the *gradient penalty* [26]. Hence we have the minimization criterion for the discriminator as

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{WD}} + \lambda_{GP}\mathcal{L}_{\text{GP}}, \tag{9}$$

where the *Wasserstein distance* criterion is

$$\begin{aligned} \mathcal{L}_{\text{WD}} = \sum_{t\geq 1} &-\mathbb{E}_{q(\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_t)} \\ &[D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t, \mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id)] + \\ &\mathbb{E}_{q(\mathbf{x}_t)p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t)}[D_\phi(\hat{\mathbf{x}}_{t-1}, \mathbf{x}_t, t, \mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id)]. \end{aligned} \tag{10}$$

with $\lambda_{\text{GP}} = 10.0$ for the *gradient penalty*:

$$\begin{aligned} \mathcal{L}_{\text{GP}} = \sum_{t\geq 1} &\mathbb{E}_{\tilde{\mathbf{x}}_{t-1}\sim\mathcal{P}(\tilde{\mathbf{x}}_{t-1})} \\ &[||\nabla_{\tilde{\mathbf{x}}_{t-1}} D_\phi(\hat{\mathbf{x}}_{t-1}, \mathbf{x}_t, t, \mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id)||_2 - 1]^2, \end{aligned} \tag{11}$$

where

$$\tilde{\mathbf{x}}_{t-1} = \alpha\mathbf{x}_{t-1} + (1-\alpha)\hat{\mathbf{x}}_{t-1}, \alpha \sim \mathcal{U}(0,1). \tag{12}$$

As defined by [22], the generator's adversarial loss is formulated against that of the discriminator as

$$\begin{aligned} \mathcal{L}_{adv} = \sum_{t\geq 1} &\mathbb{E}_{q(\mathbf{x}_t)}[D_\phi(G_\theta(\mathbf{x}_t, t, \mathbf{ms}, id), \mathbf{x}_t, t, \\ &\mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id)]. \end{aligned} \tag{13}$$

III. The Proposed System

As depicted in Fig. 2a, our proposed SVS system is composed of an acoustic model that synthesizes Mel-spectrograms from musical score inputs and a HiFiGAN neural vocoder [11] that synthesizes waveforms from estimated Mel-spectrograms. The specifications of the acoustic model can be found in Table I.

A. Acoustic Model

*1) **Transformer Encoder**:* The encoder is based on *FastSpeech2* [27] and *DiffSinger*'s encoder modules [9]. Here, we exploit Mandarin phonology to pair a note on the score with a syllable in the lyrics and decompose every syllable into a *initial-final* pair, the details of which can be found in our previous work [14]. Each *initial* and each *final* are treated as

TABLE I: Acoustic Model Specifications

Module	Hyperparameters
Transformer Decoder	Transformer-layers=4, hidden-dim=256, feed-forward size=1024, kernel-size=9
Duration Predictor	CNN-layers=3, hidden-dim=256, kernel-size=3
Diffusion Decoder	WaveNet-blocks=20, hidden-dim=256, kernel-size=3, sinusoidal-embedding-dim=512
Mel-spectrogram	sample-rate=22050Hz, hop-length=256, frame-length=1024, bins=80

a phone in the context of this work. The phone sequence and the note-length sequences are embedded, added together and passed through the Transformer stack to become the phone latent sequence.

*2) **Variance Adaptor**:* The *Variance Adaptor* contains a *duration predictor* and a *sequence-length regulator* which expands the length of the phone latent sequence from token-level to frame-level, matching that of the ground truth Mel-spectrograms, as in [5], [9], [27]. In the meantime, the note-pitch information $\mathbf{s}_{\text{pit}}$ is separately embedded, expanded by the *sequence-length regulator*, and added to the phone latent sequence to produce the frame-level hidden state sequence $\mathbf{ms} \in \mathbb{R}^{C \times L_f}$ that encompasses all the information provided by the musical score.

*3) **Diffusion Decoder**:* The architecture of the decoder $G_\theta(\mathbf{x}_t, t, \mathbf{ms}, id)$ is similar to that of [9], [18], which is essentially a stacked non-causal conditional WaveNet [1] as illustrated in Fig. 2b. With time step condition t encoded through a *sinusoidal embedding* module followed by a stack of two dense layers with *sigmoid linear unit* (SiLU) [28] non-linearity, the diffusion decoder estimates a clean Mel-spectrogram $\hat{\mathbf{x}}_{0,t}$ from the t-step noised input $\mathbf{x}_t$ conditioned on musical score information $\mathbf{ms}$. The estimated clean Mel-spectrogram is then noised through Equation 5 to complete the $p_\theta(\hat{\mathbf{x}}_{t-1}|\mathbf{x}_t)$ generation process stated by Equation 3.

B. Musical-Score-Conditioned Discriminator

The Musical-Score-Conditioned (*MSC*)-discriminator $D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t, \mathbf{s}_{\text{pho}}, \mathbf{s}_{\text{len}}, \mathbf{s}_{\text{pit}}, id)$'s backbone is 2D residual convolutional blocks (ResBlock), which are grouped and connected as depicted in Fig. 3a.

*1) **Residual Block**:* The *Residual Block* (ResBlock) illustrated in Fig 3b has two variants: the *conditional* ResBlock and the *unconditional* ResBlock. To adapt to the varying level of noise, a time step vector is projected through a dense layer and added to the residual latent features. For conditional blocks, the singer identity vector and the adaptive-interpolated encoded musical score sequence condition the residual latent features through the *Locally-Biased Modulation* (LB-mod) algorithm.

2) **Locally-Biased Modulation***:* The *Locally-Biased Modulation* (LB-mod) algorithm in the *conditional* ResBlock was inspired by [29] but differs from it in preserving the temporal dimension. *LB-mod* takes the adaptive-interpolated encoded musical score $\mathbf{ms} \in \mathbb{R}^{C_s \times W}$, the singer identity vector $\mathbf{v} \in \mathbb{R}^{C_v \times 1}$, and the latent features $\mathbf{y} \in \mathbb{R}^{C \times H \times W}$.

With two dense layers, LB-mod first projects and combines $\mathbf{ms}$ and $\mathbf{v}$ to produce the condition sequence $\mathbf{s} \in \mathbb{R}^{2C \times W}$ as in Equation 14:

$$\mathbf{s}^{2C \times W} = \text{Dense}_{\text{ms}}^{C_s \to 2C}(\mathbf{ms}) + \text{Dense}_{\text{id}}^{C_v \to 2C}(\mathbf{v}) \tag{14}$$

Subsequently, the modulation is performed as a linear-transform-and-bias operation defined in Equation 15:

$$\begin{aligned} &\textbf{LB-Mod}(\mathbf{y}^{C \times H \times W}, \mathbf{s}'^{2C \times \text{H} \times W}) := \\ &\quad \mathbf{s}'_{1:C} \odot \mathbf{y} + \mathbf{s}'_{C+1:2C}, \end{aligned} \tag{15}$$

where $\mathbf{s}'^{2C \times \text{H} \times W}$ is $\mathbf{s}^{2C \times W}$ reshaped by expanding and repeating along the second dimension.

C. Training Objective

The training objective of the acoustic model generator contains three elements: the duration loss, the reconstruction loss, and the adversarial loss.

In the training process, the ground truth phone durations $\mathbf{D}^{1 \times L_p}$ are used, while the estimated durations $\hat{\mathbf{D}}^{1 \times L_p}$ are used in the inference stage. Therefore, the *variance adaptor* returns both the encoded musical score information and the estimated duration sequence, for which a mean-square-error (MSE) is taken against the ground truth duration sequence:

$$\mathcal{L}_{\text{dur}} = \| \mathbf{D} - \hat{\mathbf{D}} \|_2 . \tag{16}$$

The diffusion decoder's Mel-spectrogram reconstruction loss is calculated as the L1-distance between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_{0,t}$:

$$\mathcal{L}_{\text{recon}} = \| \mathbf{x}_0 - \hat{\mathbf{x}}_{0,t} \|_1 . \tag{17}$$

With the adversarial loss defined by Equation 13, the acoustic model generator parameters θ are updated with gradients calculated with the weighted sum of the three objectives:

$$\mathcal{L}_G = \mathcal{L}_{\text{dur}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}. \tag{18}$$

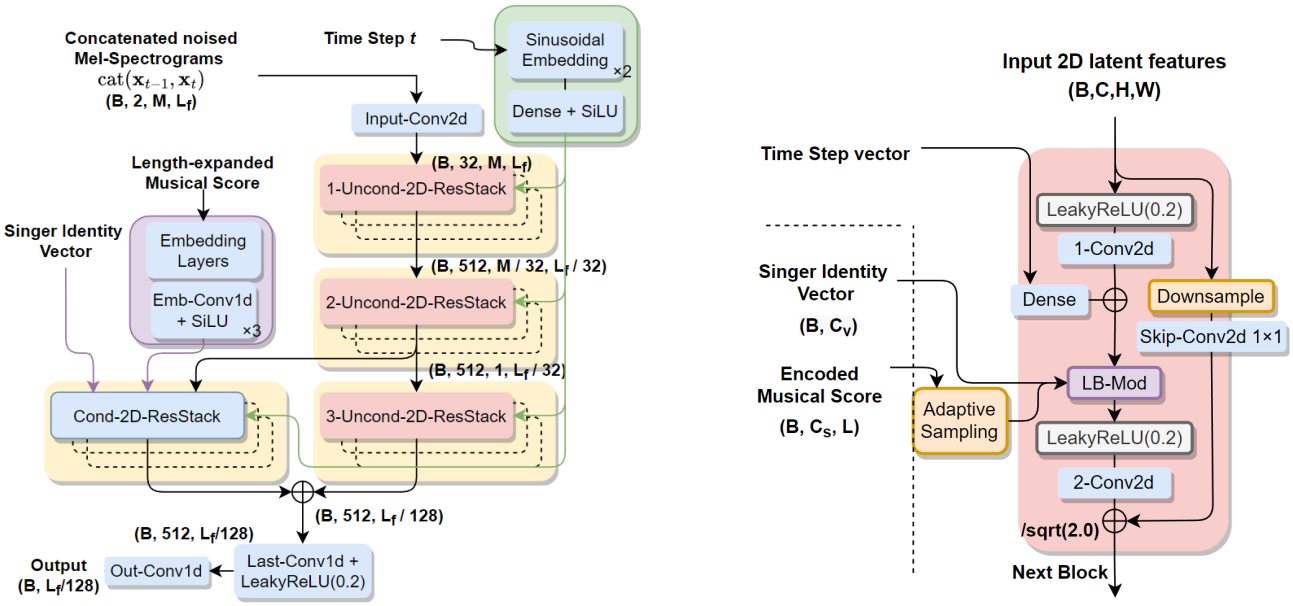
The weight for the reconstruction loss λ_{recon} and the adversarial loss λ_{adv} were both set to 1.0 for the default mixed setup. For the WGAN-only setup in the later experiment section, λ_{recon} was set to 0.

D. HiFiGAN Neural Vocoder

As the backend audio synthesizer for our SVS system, HiFiGAN [11] [2] was chosen for its compute efficiency, state-of-the-art synthesis audio quality, and robustness shown across multiple related works [8], [9], [21], [27]. For our implementation and experiments, HiFiGAN v2 was adopted with the number of initial channels modified to 256.

After the HiFiGAN was pre-trained on the Mpop600 dataset, it was fine-tuned to mitigate the discrepancies between the real and the synthesized Mel-spectrograms. In the last 35k steps of the acoustic model training, the pre-trained HiFiGAN vocoder was loaded and trained with the estimated Mel-spectrograms of the Diff-WGAN acoustic model

[2] https://github.com/jik876/hifi-gan



(a) The overall architecture of the *MSC*-discriminator.

(b) The ResBlock architecture of the *MSC*-discriminator

Fig. 3: Fig 3a: The overall architecture of the *MSC*-discriminator. Fig 3b: The internal architecture of the *Conditional*-ResBlocks used by the *MSC*-discriminator. The modules to the right of the dashed line are all block-dependent modules/operations. *Unconditional*-ResBlocks takes the same architecture, but no *Singer Identity Vector* and *Musical Score* information is provided and *LB-mod* is not performed. (H, W: the dimensions of the input downsampled from $M \times L_f$; C_s, C_v: channel dimensions of the singer identity vector and the encoded musical score respectively. Other symbols are defined the same as in Fig. 2a)

stochastically mixed into their training data according to the probability density function defined in Equation 19:

$$\Pr[\text{replace } \mathbf{x}_0 \text{ with } \hat{\mathbf{x}}_{0,t}] = (1 - \frac{t-1}{T})^2. \quad (19)$$

IV. Experiments

A. Dataset

All the experiments in this work were conducted with our lab's Mpop600 Mandarin singing voice dataset [30]. The Mpop600 dataset contains the singing voices of 2 female and 2 male singers. Each singer contributed 3 hours of singing voice audio consisting of 150 different song excerpts. In total, the whole Mpop600 dataset is 12-hour-long with 600 song excerpts. All the songs in this dataset were pop-music sung in Mandarin and recorded without background instruments. The audio and phone-transcribed lyrics were forced-aligned to frame-level by the open-source Speech-Aligner[3], the details of which can be found in our previous work [14], [30]. For all the experiments, the audio recordings were down-sampled from 96 kHz, 24-bits to 22.05 kHz, 16-bits.

For evaluation, 2 song excerpts from each singer's 150 song excerpts were reserved and not seen by the model in the training process. To train the models with mini-batches, every song excerpt was segmented into smaller audio samples with audio durations of 6 to 12 seconds.

[3]https://github.com/open-speech/speech-aligner

B. Training Setup

Experiments were conducted with audio samples produced through the following setups:

- **Reference**: ground truth Mpop600 audio samples sung by human singers down-sampled to 22.05 kHz.
- **Resynthesized**: ground truth audio samples converted to Mel-spectrograms and re-synthesized by the HiFiGAN vocoder at 22.05 kHz sample rate.
- **model-FFT**: an end-to-end SVS system with a *Feed-forward Transformer* backbone as the acoustic model decoder trained on an L1 Mel-spectrogram estimation loss.
- **model-Diff-L1** ($T = 4$): the proposed system with the DDPM acoustic model trained exclusively on the L1 Mel-spectrogram reconstruction loss without WGAN.
- **model-Diff-Mixed** ($T = 4$): the proposed system with the DDPM acoustic model trained with both the L1 Mel-spectrogram reconstruction loss and the WGAN adversarial loss.
- **model-Diff-WGAN** ($T = 4$): the proposed system with the DDPM acoustic model trained exclusively by WGAN without the L1 Mel-spectrogram reconstruction loss.

All the end-to-end synthesis models utilized the HiFiGAN vocoder for waveform generation and Mel-specrtograms as the acoustic features. The experimented acoustic models were trained by an *AdamW* [31] optimizer with L2 weight decay of 10^{-6} while the the HiFiGAN vocoder was pre-trained with

System	MS-SSIM ↑	MCD ↓	F0-RMSE ↓	F0-Corr ↑
Resynthesized	0.985	2.613	0.273	0.936
FFT	0.756	8.206	0.628	0.751
Diff-L1	**0.899**	**5.702**	**0.411**	**0.882**
Diff-Mixed	0.879	6.348	0.463	0.881
Diff-WGAN	0.886	6.130	0.438	0.878

TABLE II: Objective evaluation results of different systems.

an *Adam* [32] optimizer without weight decay.

For model-FFT, *AdamW* hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\mathrm{lr} = 10^{-4}$ were applied with the learning rate scheduling proposed for *Transformer*s in [33]. For all the diffusion-based models, the *AdamW* hyperparameters were $\beta_1 = 0.5$, $\beta_2 = 0.9$. Since model-Diff-L1 was trained solely on reconstruction objectives, its learning rate was exponentially decay by a rate of 0.999 at the start of every epoch with initial learning rate $\mathrm{lr} = 10^{-4}$. For the two Diff-WGAN acoustic models (model-Diff-Mixed, model-Diff-WGAN), the learning rate was 10^{-4} for both the generator and the discriminator, and no learning rate scheduling was applied.

All the acoustic models were trained to 210k steps, taking 16 hours for model-FFT, 20 hours for model-Diff-L1, and 48 hours for the two Diff-WGAN model since the discriminator and the acoustic model were updated with a $2:1$ ratio as WGAN requires the discriminator to converge faster than the generator [22], [26]. These training sessions in the experiment were completed on one RTX 3090 graphics card with a batch size of 8.

C. Objective Evaluation

For perceptual similarity, we employed two objective metrics: *Mel-Cepstral-Distortion* (MCD) commonly adopted for audio signals and the *Multi-Scale Mean Structural Similarity* (MS-SSIM) [34] metric. Here, MS-SSIM was applied to the Mel-spectrograms of the synthesized and ground truth signal to evaluate their structural similarity on the spectral domain as images [35].

To evaluate the fundamental frequency (F0) fidelity of the SVS system on the test dataset, we employed the *root-mean-square-error* (RMSE) and *Pearson-correlation* (Corr) metrics on the F0 sequences extracted by the *WORLD* vocoder [4] [4]. To eliminate the effects of rhythmic swings, *Dynamic Time Warping* (DTW) was first performed between the ground truth signal and the compared signal with an open-source fast-DTW tool [5]. For the musical context of this work, the F0-RMSE values were calculated on the *semitone* scale.

The objective evaluation results of the compared systems is presented in Table II.

D. Mean-Opinion-Score evaluation

To understand each system's synthesis quality perceived by the human listener, 10 participants were asked to rate each sample out of $\{1, 2, 3, 4, 5\}$, where *1* indicates *worst* and *5* indicates *best* in terms of *"naturalness concerning human singers"*. Two samples from each of the four singers were included in the MOS test. The MOS data is presented in Table III with their 95% confidence intervals.

System	MOS Score $(1 \sim 5)$
Reference	4.76 ± 0.099
Re-synthesized	4.15 ± 0.243
model-FFT	2.55 ± 0.447
model-Diff-L1	$\mathbf{3.79 \pm 0.242}$
model-Diff-Mixed	3.49 ± 0.481
model-Diff-WGAN	2.88 ± 0.279

TABLE III: Mean Opinion Score (MOS) evaluation results of different systems collected on 10 human participants.

E. Preference Test on Musical Expressiveness

To particularly examine the proposed Diff-WGAN architecture's effects on musical expressiveness, model-Diff-L1 and model-Diff-WGAN were put in a preference test with 20 participants. Four extra audio segments were sampled from the validation set of the two skilled singers (one male and one female) who exhibit more complex variations musically. The audio samples were formulated into four questions. In each question, the human participants had to choose the sample they deemed *more musically expressive disregarding pronunciation accuracy or audio quality* or choose the *No Preference* option if they deemed the two on par. The result of the preference test is shown in Fig 4.

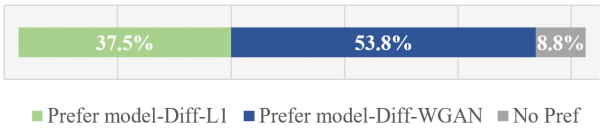


Fig. 4: Musical expressiveness preference test results on 20 participants.

V. Discussion and Analysis

A. Discussion on Quantitative Evaluation Results

In Table III's MOS test results, model-Diff-L1 received the highest score. Looking at the measurements presented in Table II, model-Diff-L1's advantage can likely be attributed to its reconstructive L1 objective. Since its target was only to best-estimate the ground truth Mel-spectrograms in the L1 distance, model-Diff-L1 naturally has a low overall distortion to the ground truth, as evident by the objective measurements. This attribute means model-Diff-L1 robustly produces the high-energy spectral features that mimic those produced by human singers and rids the synthesized Mel-spectrograms of

[4] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder
[5] https://github.com/slaypni/fastdtw

significant artifacts that greatly impact the perceived quality of an audio segment in the singing voice context.

In comparison, model-Diff-Mixed and model-Diff-WGAN's objectives were not to estimate the ground truth Mel-spectrograms but their generation distributions; therefore, the generator was enforced to also reproduce the variations and inconsistencies of human singers. However, in the context of sung music, human listeners give audio samples with any artifact or error in pitch or pronunciation significantly lower scores regardless of the rest of their properties. Hence, the less consistent Diff-WGAN-based models received lower scores in both the objective measurements and in the MOS test.

Nevertheless, model-Diff-WGAN crucially beats model-Diff-L1 in the musical expressiveness preference test. While the best strategy for model-Diff-L1 to optimize for reconstruction objectives was to over-smooth these variations, model-Diff-WGAN optimized towards the data distribution estimated by the acoustic model discriminator, which injects variations and details not just acoustically but also on meta-features as musical expression. This observation is further investigated in the following qualitative study.

B. Qualitative Study on Generated Mel-Spectrograms

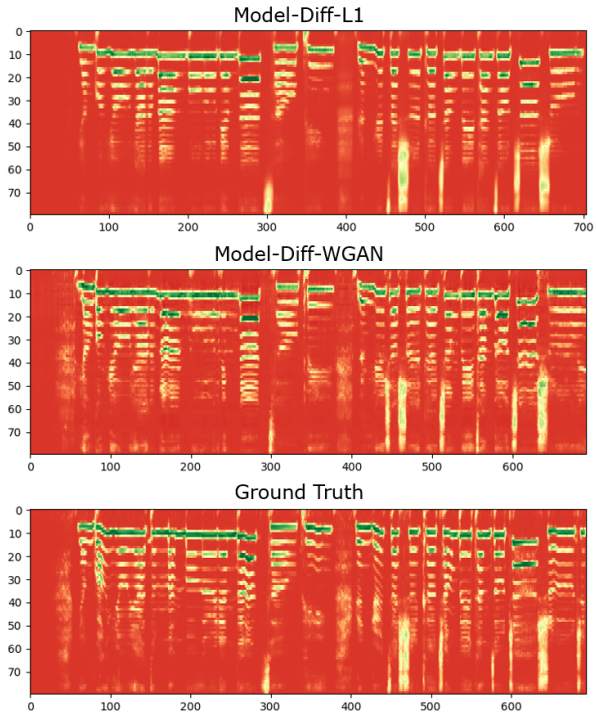


Fig. 5: A set of Mel-spectrogram samples from model-Diff-L1, model-Diff-WGAN, and extracted directly from ground truth human singer's audio in the validation dataset.

The synthesis characteristics of acoustic models with or without WGAN can be understood by observing their resulting Mel-spectrograms. As illustrated in Fig. 5, model-Diff-L1 trained only with reconstruction loss and model-Diff-WGAN trained only with WGAN produced Mel-spectrograms with distinct styles. Model-Diff-WGAN's synthesized Mel-spectrogram contains high-frequency components, noises during silent segments, and variations of energy reminiscent of the ground truth Mel-spectrogram.

Contrarily, the high-frequency details on the model-Diff-L1's spectrogram are not just less prominent, but those that are present are also less granular than the high-frequency details on model-Diff-WGAN's or the ground truth's Mel-spectrogram. Nevertheless, when our attention focuses on the potent high-energy components in the lower frequency bins, model-Diff-L1's Mel-spectrogram shows a more consistent and smooth change in energy, and the fundamental frequency contour exhibits clear-cut discontinuities at inter-syllable unvoiced consonants. These properties combined mean the resulting signal does not have unnatural artifacts and gives the perception of properly accentuated syllable pronunciations.

Based on these observations, it can be concluded that WGAN has indeed prevented the acoustic model from over-smoothing, but the variations it introduced also made the acoustic model less consistent and prone to error. As human listeners are intolerable to errors in music, further investigations have to be made to leverage the correctness enforced by the reconstruction objective and the spectral details and expressiveness brought about by the WGAN.

C. Remark on Training Stability and Convergence

One of the key advantages of the proposed combination of DDPM and WGAN is its stability. In the training process, the Diff-WGAN acoustic models' *Wasserstein* discriminator loss decreased monotonically, and no mode collapse has been encountered in the experimented configurations even for model-Diff-WGAN, in which no reconstruction loss was employed as constraint or guidance.

VI. Conclusions

This work proposed an acoustic model based on a combined architecture of diffusion denoising probabilistic model (DDPM) and *Wasserstein* generative adversarial network (WGAN). To exploit the singing voice synthesis (SVS) formulation, the discriminator was designed to be conditioned not only on the synthesis features but also on the musical score and singer identity. This proposed acoustic model architecture was implemented with an integrated HiFiGAN vocoder to form a multi-singer end-to-end singing voice synthesis system trained on the Mpop600 Mandarin singing voice dataset. Although the proposed architecture was not fully exploited to comprehensively outperform the baseline diffusion model, the preference test and the qualitative study suggested that the addition of WGAN had fulfilled its purpose of prompting musical expressiveness and enforcing high-frequency acoustic details. Moreover, this adversarial acoustic model was shown to converge without the reconstruction objective as guidance, proving the convergency and stability of the proposed Diff-WGAN architecture.

VII. Acknowledgment

This research is supported by the Ministry of Science and Technology of Taiwan under Grant No. 109-2221-E-007-094-MY3.

References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, p. 125, ISCA, 2016.

[2] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7237–7241, 2022.

[3] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, "Adversarially Trained End-to-End Korean Singing Voice Synthesis System," in *Proc. Interspeech 2019*, pp. 2588–2592, 2019.

[4] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[5] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System," in *Proc. Interspeech 2020*, pp. 1306–1310, 2020.

[6] X. Zhuang, T. Jiang, S.-Y. Chou, B. Wu, P. Hu, and S. Lui, "Litesing: Towards fast, lightweight and expressive singing voice synthesis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7078–7082, 2021.

[7] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7229–7233, 2020.

[8] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *ArXiv*, vol. abs/2009.01776, 2020.

[9] J. Liu, C. Li, Y. Ren, F. Chen, P. Liu, and Z. Zhao, "Diffsinger: Diffusion acoustic model for singing voice synthesis," *arXiv*, vol. abs/2105.02446, 2021.

[10] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2020.

[11] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[12] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, "Ddsp-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation," *Proc. International Society for Music Information Retrieval*, 2022.

[13] J. Kim, H. Choi, J. Park, M. Hahn, S. Kim, and J.-J. Kim, "Korean Singing Voice Synthesis Based on an LSTM Recurrent Neural Network," in *Proc. Interspeech 2018*, pp. 1551–1555, 2018.

[14] F.-R. Yang, Y.-P. Cho, Y.-H. Yang, D.-Y. Wu, S.-H. Wu, and Y.-W. Liu, "Mandarin singing voice synthesis with a phonology-based duration model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1975–1981, 2021.

[15] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6955–6959, 2019.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[17] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[18] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations*, 2021.

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

[20] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *International Conference on Learning Representations*, 2022.

[21] S. Liu, D. Su, and D. Yu, "Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans," *ArXiv*, vol. abs/2201.11972, 2022.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *ArXiv*, vol. abs/1701.07875, 2017.

[23] J. Wu, Z. Huang, J. Thoma, D. Acharya, and L. Van Gool, "Wasserstein divergence for gans," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[24] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[25] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2020.

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[27] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2021.

[28] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks : the official journal of the International Neural Network Society*, vol. 107, pp. 3–11, 2018.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[30] C.-C. Chu, F.-R. Yang, Y.-J. Lee, Y.-W. Liu, and S.-H. Wu, "MPop600: A Mandarin popular song database with aligned audio, lyrics, and musical scores for singing voice synthesis," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1647–1652, 2020.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[34] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003*, vol. 2, pp. 1398–1402 Vol.2, 2003.

[35] C. Gan, X. Wang, M. Zhu, and X. Yu, "Audio quality evaluation using frequency structural similarity measure," in *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2011)*, pp. 299–303, 2011.