IEEE Open Journal of Signal Processing

Received 12 December 2023; revised 16 March 2024; accepted 28 March 2024. Date of publication 8 April 2024; date of current version 25 April 2024. The review of this article was arranged by Associate Editor E. Variani.

Digital Object Identifier 10.1109/OJSP.2024.3386495

Adversarial Training of Denoising Diffusion Model Using Dual Discriminators for High-Fidelity Multi-Speaker TTS

MYEONGJIN KO, EUIYEON KIM, AND YONG-HOON CHOI (Member, IEEE)

School of Robotics, Kwangwoon University, Seoul 01897, South Korea

CORRESPONDING AUTHOR: YONG-HOON CHOI (e-mail: yhchoi@kw.ac.kr).

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government, MSIT, under Grant 2021R1F1A1064080.

ABSTRACT The diffusion model is capable of generating high-quality data through a probabilistic approach. However, it suffers from the drawback of slow generation speed due to its requirement for many time steps. To address this limitation, recent models such as denoising diffusion implicit models (DDIM) focus on sample generation without explicitly modeling the entire probability distribution, while models like denoising diffusion generative adversarial networks (GAN) combine diffusion processes with GANs. In the field of speech synthesis, a recent diffusion speech synthesis model called DiffGAN-TTS, which utilizes the structure of GANs, has been introduced and demonstrates superior performance in both speech quality and generation speed. In this paper, to further enhance the performance of DiffGAN-TTS, we propose a speech synthesis model with two discriminators: a diffusion discriminator to learn the distribution of the reverse process, and a spectrogram discriminator to learn the distribution of the generated data. Objective metrics such as the structural similarity index measure (SSIM), mel-cepstral distortion (MCD), F0 root mean squared error (F0-RMSE), phoneme error rate (PER), word error rate (WER), as well as subjective metrics like mean opinion score (MOS), are used to evaluate the performance of the proposed model. The evaluation results demonstrate that our model matches or exceeds recent state-of-the-art models like FastSpeech 2 and DiffGAN-TTS across various metrics. Our code and audio samples are available on GitHub.

INDEX TERMS Denoising diffusion model, generative adversarial network, mel-spectrogram discriminator, speech synthesis, text-to-speech.

I. INTRODUCTION

Generative models [1], [2], [3], [4], [5] are artificial intelligence frameworks capable of producing new data types beyond those seen during training. Applied across domains such as language processing, creative arts, image creation, editing, and speech synthesis, they are pivotal in technological advancement. The variational autoencoder (VAE) [1], one of the earliest and most representative generative models, encodes input data into a latent space before reconstructing it back into data. The VAE has garnered significant attention and remains a prominent model in the field. Meanwhile, the generative adversarial network (GAN) [2], still the most widely recognized among generative models, generates high-quality data indistinguishable from real data through an adversarial training approach involving a discriminator and a generator. Another type of generative model, the flow-based generative model [3], explicitly learns the probability distribution of the latent space and uses inverse transformations to generate data. However, these representative models still face limitations such as unstable training, limited diversity in generated results due to mode collapse, and loss function dependence on model architecture.

The diffusion model [4], inspired by non-equilibrium thermodynamics in physics, reduces noise while preserving important details by adding and removing noise in the data, based on the Markov chain principle where each variable's value depends on the previous time step. Trained through a predefined procedure, this model learns a latent space representation to restore data, enabling new data point generation directly from random noise or the encoded information in

the latent space. Unlike traditional generative models, the diffusion model boasts high training stability and does not compress data, allowing the latent space to maintain the same high dimensionality as the original data. In computer vision, numerous models have been developed based on the characteristics of the diffusion model, beginning with denoising diffusion probabilistic models (DDPM) [5].

A speech synthesis system, also known as a text-to-speech (TTS) system, transforms written scripts into spoken words. Early TTS systems relied on concatenative synthesis, including unit selection [6], stitching together phoneme sounds, and statistical parametric methods using hidden Markov models (HMM) [7]. These approaches, however, struggle to produce lifelike speech. The introduction of neural network models has significantly enhanced the quality and naturalness of speech synthesis.

One notable speech synthesis model based on neural networks is Tacotron [8]. It utilizes a sequence-to-sequence architecture [9] and attention mechanism [10] to generate mel-spectrograms autoregressively. These are then transformed into linear spectrograms, and the audio waveform is produced using the Griffin-Lim vocoder [11]. Tacotron 2 [12] addresses the complex hyperparameter adjustment issues faced by the previous model and improves audio quality by integrating the WaveNet vocoder [13]. However, it still contends with slow generation speeds and potential error accumulation due to its autoregressive design. Recently, GAN models applied to vocoders have significantly improved generation speed and audio quality. FastSpeech 2 [14] and FastPitch [15], leveraging the Transformer architecture, produce mel-spectrograms non-autoregressively, addressing the slow speeds and error issues encountered by Tacotron models. FastSpeech 2 incorporates a variance adaptor for speech characteristics control.

The TTS models are evolving to quickly generate natural and high-quality speech. Diffusion models are considered suitable for speech synthesis due to their capability to produce high-quality data. However, conventional diffusion models are hindered by slow generation speeds, rendering them unsuitable for real-time applications commonly used in TTS systems. The demand for faster synthesis is prevalent not only in computer vision but also in speech synthesis, spurring ongoing research to enhance the generation speed of diffusion models [16], [17]. A prominent example is the denoising diffusion implicit model (DDIM) [16], which builds upon the structure of the DDPM [5] to improve data generation speed. DDIM adopts a non-Markovian approach, diverging from the Markov chain process used in original diffusion models. In its training phase, DDIM models the reverse process akin to DDPM, where the process relies on incrementally increasing time steps. However, during generation, it employs time steps as conditional inputs, enabling the model to bypass intermediate stages, moving from the initial noise $\mathbf{x}_T$ directly to an arbitrary time step t state, denoted as $\mathbf{x}_t$, through accelerated sampling. This method significantly reduces the number of required time steps compared to DDPM. The denoising diffusion GAN model [17] integrates the GAN structure into diffusion models. The noise added or removed in this process adheres to a Gaussian distribution. In [17], the slow generation speed of traditional diffusion models is linked to the use of a Gaussian distribution for sampling data during the reverse process. To tackle this, [17] introduces multimodal distributions for reverse diffusion, hastening the generation speed. Moreover, this model utilizes an adversarial training approach characteristic of GANs, enabling it to learn the necessary noise distribution for the reverse process.

As the generation speed of diffusion models has improved, research is ongoing to apply these models to acoustic models and vocoders. A prominent example is Diff-TTS [18], an acoustic model that employs the accelerated sampling technique used in DDIM. Diff-TTS exhibits superior audio quality compared to Tacotron 2 or Glow-TTS [19] when generating speech without accelerated sampling. However, employing accelerated sampling increases generation speed at the cost of reduced audio quality. DiffGAN-TTS [20] is a cutting-edge acoustic model based on the denoising diffusion GAN structure. It facilitates text-to-speech synthesis for multiple speakers, with its generator derived from FastSpeech 2, allowing for controlled speech generation. However, DiffGAN-TTS may face challenges in effectively learning detailed elements as it relies on a single discriminator to learn both the multimodal distribution for the reverse process and the characteristics of voices from multiple speakers. Huang et al. [21] analyze the trade-offs between sample diversity, quality, and computational efficiency in speech synthesis using combined GAN and DDPM models. They introduce FastDiff 2, which merges GANs and DDPMs into two variations: DiffGAN, utilizing a conditional GAN-based denoising process for stable, large-step reverse processing; and GANDiff, a GAN incorporating diffusion iterations for enhanced sample diversity. Both variations showcase superior speech synthesis with high quality and diversity through an efficient four-step sampling process. Deng et al. [22] introduce MixGAN-TTS, a non-autoregressive speech synthesis model that merges and enhances features from PortaSpeech [23] and DiffGAN-TTS [20]. This model addresses the ambiguity in phoneme boundaries, incorporates a mixed alignment mechanism and pitch and energy predictors for better audio variance handling. Unlike traditional diffusion models that suffer from slow real-time performance due to the Gaussian function's limitations, MixGAN-TTS employs GAN for modeling the denoising distribution, enabling the generation of high-quality audio efficiently with fewer denoising steps.

In this paper, we propose an acoustic model named SpecDiff-GAN that integrates diffusion models with GANs. This model adopts the generator from DiffGAN-TTS [20] and features two discriminators: a diffusion discriminator and a spectrogram discriminator. This architecture enables separate and independent training of the features necessary for the reverse process and the distinct characteristics of various speakers' voices. To assess the performance of SpecDiff-GAN, we conduct experiments in multi-speaker speech synthesis. Objective evaluations are performed to determine how accurately

the synthesized speech mirrors speaker characteristics and the overall speech quality. These evaluations include measurements of the structural similarity index measure (SSIM) [24], mel-cepstral distortion (MCD) [25], F0 root mean squared error (F0-RMSE), phoneme error rate (PER), word error rate (WER) [26], and real-time factor (RTF). Additionally, subjective evaluations are carried out using comparative mean opinion score (CMOS) and similarity mean opinion score (SMOS).

The remainder of this paper is organized as follows: Section II explains the operational principles of the diffusion model. Section III provides a detailed description of the proposed model. Sections IV and V describe the experimental setup and present the experimental results of the proposed model, respectively. Finally, Section VI concludes the study and discusses future research directions.

II. BACKGROUND

In TTS, diffusion models follow a *diffusion process* where Gaussian noise is gradually added to the (mel)-spectrogram through Markov chain transitions, converting it into a latent vector. They also employ a *reverse process* (also known as the *denoising process*) that removes noise from the latent vector and reconstructs the spectrogram. Let $\mathbf{x}_t \in \mathbb{R}^L$ for $t = 0, 1, \ldots, T$ be a sequence of corrupted spectrograms with the same dimension, where t is the index for diffusion time step. Each diffusion process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right), \tag{1}$$

where $\beta_t \in (0, 1)$ is a hyperparameter predefined ahead of model training. The whole diffusion process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is defined by Markov chain transition:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t\geq 1} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{2}$$

The reverse process is the inverse of the diffusion process and serves as the procedure for generating spectrogram from Gaussian noise. Each denoising process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ follows:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right), \tag{3}$$

where $\mu_\theta(\mathbf{x}_t, \mathrm{t})$ and σ_t^2 are the mean and variance for the denoising model. The whole reverse process parameterized with θ is defined by:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t\geq 1} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t). \tag{4}$$

The latent vector is gradually restored to a spectrogram through the reverse transitions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The goal of training is to maximize the likelihood $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}$, but since it is intractable, the model is trained to maximize the evidence lower bound (ELBO, $\mathcal{L} \leq \log p_\theta(\mathbf{x}_0)$) instead.

The key assumption in the diffusion model is that the noise levels at each step are small, which allows the diffusion process to be stable and the model to be tractable. The number of denoising steps T is often assumed to be in the order of hundreds to thousands. Therefore, the parameters $\beta_1, \beta_2, \ldots, \beta_T$, which determine the amount of noise to be added during diffusion, are set to be small. As a result, the size of the diffusion time steps T increases, leading to a significant time requirement for spectrogram reconstruction.

To reduce the number of denoising steps, Xiao et al. [17] proposed modeling the denoising distribution with a complex multimodal distribution. They introduced the denoising diffusion GAN, which models each denoising step using a multimodal conditional GAN. This approach significantly reduced the denoising time steps to a level feasible for real-time applications. However, due to the reduced number of time steps, the quality of the generated data may be lower compared to the DDPM, with some features potentially not being adequately represented during noise removal from the latent space and data generation.

Research [17] and [20] have combined diffusion models with GAN architectures to significantly reduce the timesteps required in diffusion and denoising processes, thereby enabling high-quality speech generation. The discriminator in these models must concurrently learn the characteristics of data changes during the denoising process and the features of the mel-spectrogram, crucial for training the acoustic model. In the case of DiffGAN-TTS [20], a joint conditional and unconditional (JCU) discriminator, akin to that in VocGAN, is used to simultaneously learn these two feature types. Conversely, high-performance voice synthesis models like HiFi-GAN employ a multi-discriminator structure, facilitating the learning of diverse vocal characteristics by having dedicated discriminators for each feature. Drawing inspiration from this structure, our proposed model is designed with distinct discriminators for learning features in the denoising process as well as for other voice-related characteristics, such as the mel-spectrogram or speaker traits, thus enhancing the model's performance.

III. MODEL DEVELOPMENT

The overall architecture of the proposed SpecDiff-GAN in this paper consists of one generator and two discriminators, as shown in Fig. 1. The DiffGAN-TTS architecture [20] serves as the underlying base model for generating high-quality synthesized speech.

A. THE GENERATOR OF SPECDIFF-GAN

The generator structure of SpecDiff-GAN, depicted in Fig. 2, is based on the generator structure of DiffGAN-TTS. This generator is a modified version of the FastSpeech 2 [14] architecture, comprising four transformer encoders, one variance adaptor, and one diffusion decoder. The transformer encoder processes the phoneme embedding sequence as input, converting it into a hidden sequence, akin to the FastSpeech 2 approach. The variance adaptor, consisting of duration, pitch, and energy predictors, analyzes the hidden sequence from the encoder to predict appropriate speech length, pitch, and energy, facilitating the generation of natural-sounding speech.

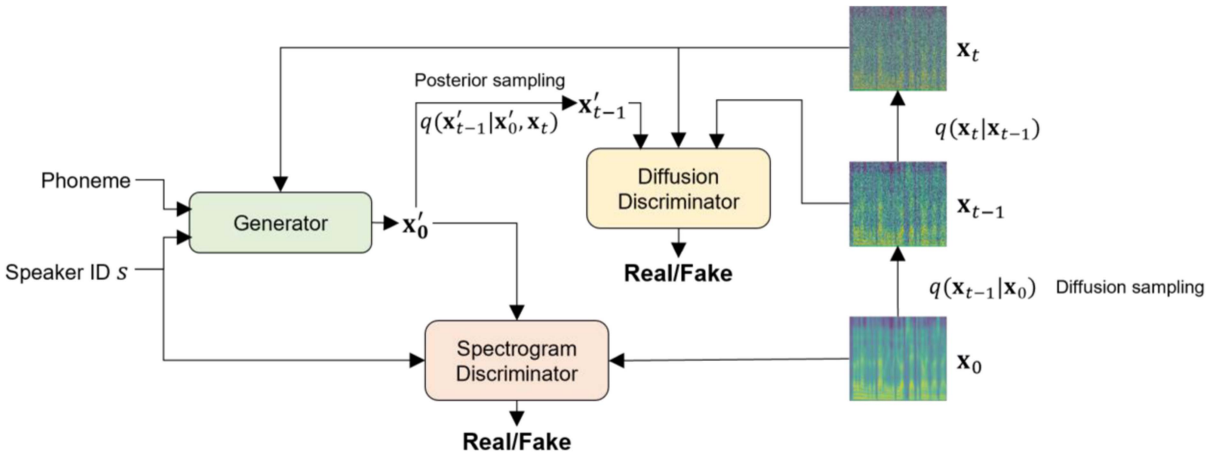

FIGURE 1. SpecDiffGAN-TTS architecture.

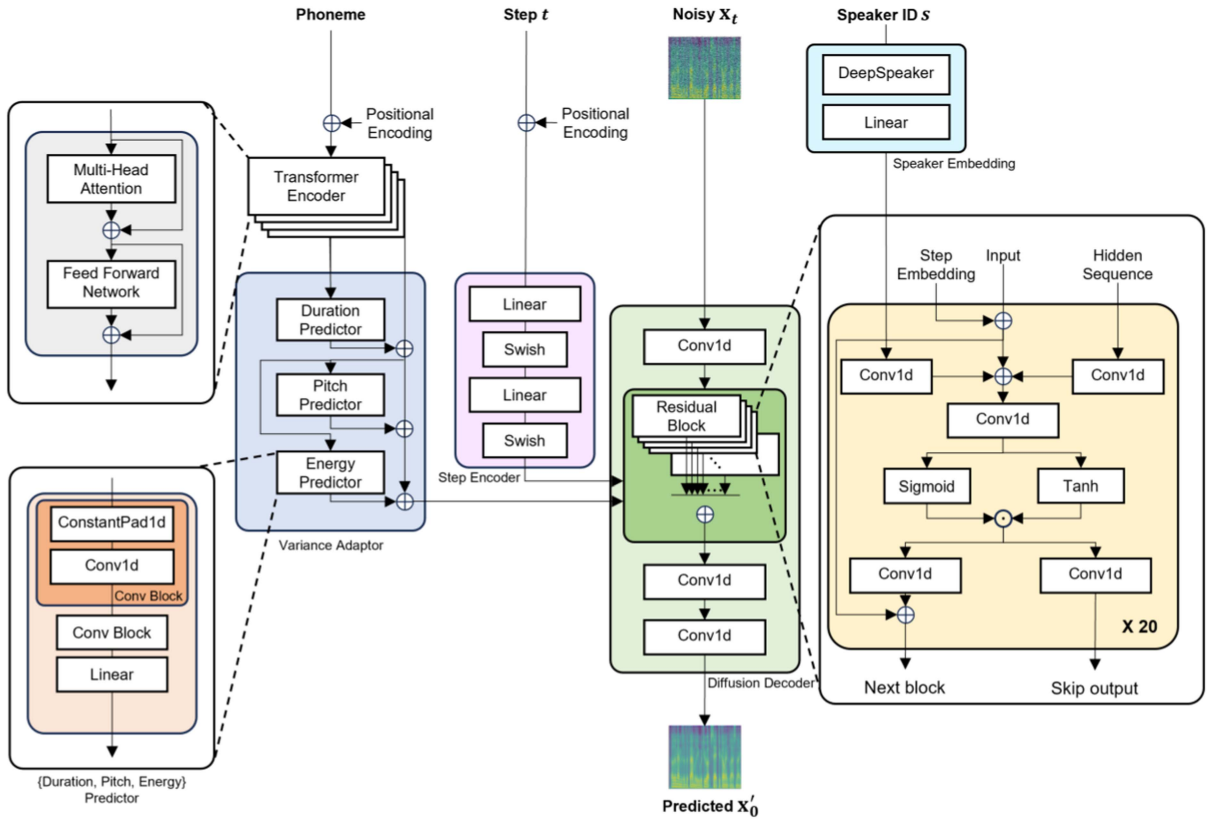

FIGURE 2. Generator structure of SpecDiff-GAN.

The decoder is designed by replacing the decoder of the FastSpeech 2 with a diffusion decoder.

The diffusion decoder in SpecDiff-GAN is inspired by the structure of WaveNet [13], an autoregressive vocoder model. However, it differs from WaveNet in certain aspects. While the WaveNet vocoder utilizes dilated causal convolutions to generate waveforms, the diffusion decoder in SpecDiff-GAN employs conventional convolution layers, as mel-spectrograms do not require as extensive receptive fields as waveforms. The diffusion decoder of the proposed model accepts speaker embeddings, which capture each speaker's unique characteristics, as conditional inputs. This feature enables the generation of distinct voices for different speakers. Deep Speaker [27] is used to extract these speaker embeddings. Additionally, sinusoidal positional encoding [10] is applied to represent the time step t at each stage, allowing for the accurate modeling of appropriate distributions for each time step.

The feature map generated by the transformer encoder and variance adaptor is passed through a 1×1 convolution layer and then input into each residual block of the diffusion

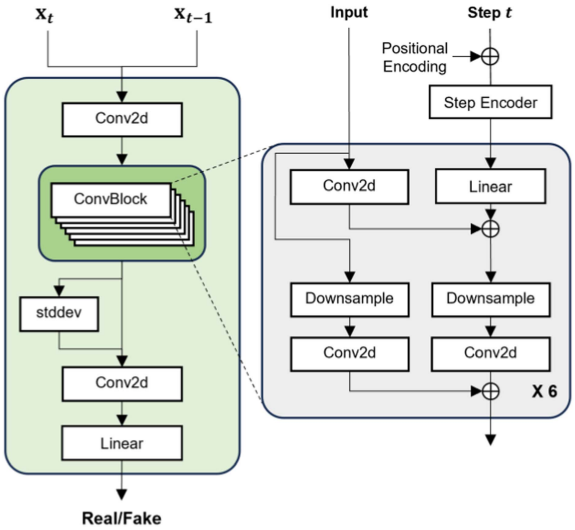

FIGURE 3. Diffusion discriminator of SpecDiff-GAN.

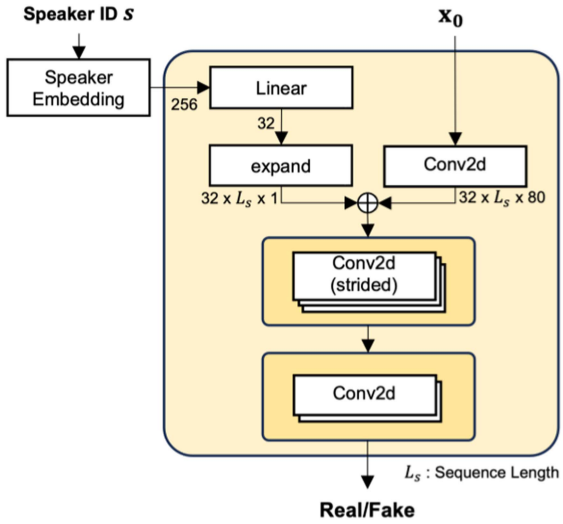

FIGURE 4. Spectrogram discriminator of SpecDiff-GAN.

decoder, along with the speaker embedding. The hidden features output from each residual block are combined using skip connections and then fed through two 1×1 convolution layers to generate the mel-spectrogram.

In summary, the acoustic generator that produces the mel-spectrogram $\mathbf{x}'_0$ can be modeled as $G(\mathbf{x}_t, \mathbf{y}, s, t)$, where $\mathbf{x}_t$ is the corrupted mel-spectrogram, $\mathbf{y}$ is the phoneme sequence input, s represents the speaker ID, and t denotes the diffusion step index.

B. THE DIFFUSION DISCRIMINATOR

The diffusion discriminator structure of the proposed model, depicted in Fig. 3, is based on the discriminator structure of progressive growing of GANs (ProGAN) [28], with a time-dependent modification. The diffusion discriminator consists of multiple downsampling convolution blocks, each comprising downsampling layers and 2D convolution layers. Timestep embeddings are conditionally applied to the input values for downsampling. The features extracted from both conditioned and unconditioned downsampling are then combined and fed into the subsequent block. This process is repeated across six blocks.

Mel-spectrograms $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ are concatenated to overlap and used as inputs to the diffusion discriminator, which iteratively compresses the data and extracts features to determine whether the reverse process from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ is performed effectively. In ProGAN, to address the issue of mode collapsing that can often occur in GAN training, the minibatch discrimination technique [29] was employed, which calculates the closeness among the data within a batch. This technique is also applied in the diffusion discriminator to enhance the stability of the training process.

C. THE SPECTROGRAM DISCRIMINATOR

The structure of the spectrogram discriminator in the proposed model, as shown in Fig. 4, is inspired by the multi-resolution spectrogram discriminator of UnivNet [30], a GAN-based vocoder model. While UnivNet employs multiple sets of short-time Fourier transform (STFT) parameters applied to mel-spectrograms for audio waveform generation, our focus in this paper is on using mel-spectrograms as an acoustic model for mel-spectrogram generation, rather than waveforms. Consequently, the multi-resolution input is not utilized; instead, only the hierarchical structure of UnivNet's discriminator is adopted.

In multi-speaker speech synthesis models, learning distinctive features for individual speakers is crucial. In our approach, speaker embeddings are used as conditional inputs to the discriminator. The speaker embedding passes through a linear layer, while the mel-spectrogram is processed through a 2D convolution layer. This is followed by broadcasting and addition operations. The spectrogram discriminator then repeats the convolution layers depicted in Fig. 4, enabling it to discern the unique characteristics of each speaker's voice within the mel-spectrograms.

In [20], the primary focus of a single diffusion discriminator is on the diffusion properties of the input spectrogram. While adding a spectrogram discriminator introduces additional costs, including increased model complexity and training time, it offers several advantages:

- Enhanced modeling of spectral features: While the diffusion discriminator concentrates on the temporal evolution of the spectrogram during the diffusion process, a spectrogram discriminator can specifically focus on capturing spectral features, including frequency components and patterns. These may be missed or inadequately represented by the diffusion discriminator alone.
- Improved frequency and time detail handling: A spectrogram discriminator excels at capturing fine-grained details in both frequency and time domains. This is crucial for handling high-frequency components and capturing subtle temporal variations in the input spectrogram. The addition of a spectrogram discriminator

allows the model to better capture nuances in the spectral content.

- Diversity in discriminative aspects: Utilizing both a diffusion and spectrogram discriminator allows the model to consider different aspects. The diffusion discriminator focuses on dynamic evolution, while the spectrogram discriminator provides insights into detailed spectral characteristics, fostering a more comprehensive understanding.
- Enhanced discrimination in both time and frequency domains: The spectrogram discriminator treats the spectrogram as a 2D image, allowing it to exploit correlations among different components in both the time and frequency domains. This enables the model to discriminate not only based on temporal changes but also on specific frequency components, providing a more comprehensive discriminative capability.

D. LOSS FUNCTIONS

In this paper, the least squares GAN (LSGAN) [31] loss function is employed to train two discriminators. This loss function is utilized to prevent gradient vanishing and has already been proven to be effective when applied to the field of audio and speech synthesis. The diffusion discriminator, denoted as $D_d(\mathbf{x}_{t-1},\ \mathbf{x}_t,\ t)$, is trained to minimize the loss:

$$\mathcal{L}_{diff} = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t)q(\mathbf{x}_{t-1}|\mathbf{x}_t)} \left[(D_d\left(\mathbf{x}_{t-1},\ \mathbf{x}_t,\ t\right) - 1)^2 \right] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \left[\left(D_d\left(\mathbf{x}'_{t-1},\ \mathbf{x}_t,\ t\right) \right)^2 \right], \tag{5}$$

where t denotes diffusion time step index. Equation (5) is the same loss function used in [20], but the speaker ID is not used as an argument. The spectrogram discriminator, denoted as $D_s(\mathbf{x}_0,\ s)$, is trained to minimize the loss:

$$\mathcal{L}_{spec} = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}(\mathbf{x}_0)} \left[(D_s\left(\mathbf{x}_0,\ s\right) - 1)^2 \right] + \mathbb{E}_{\mathbf{x}'_0 \sim p_\theta(\mathbf{x}_{0:T})} \left[\left(D_s\left(\mathbf{x}'_0,\ s\right) \right)^2 \right], \tag{6}$$

where s denotes the speaker ID.

In the SpecDiff-GAN generator, three loss functions are used. The feature matching loss is employed to ensure that the distribution of the generated data matches the distribution of the real data and to prevent the discriminator from overfitting. Feature matching loss $\mathcal{L}_{fm}$ is calculated by summing the l_1 distance between all discriminator feature maps of the generated and real data: where $D_d^i(\cdot)$ represents the i-th hidden layer of the diffusion discriminator, and $D_s^i(\cdot)$ represents the i-th hidden layer of the spectrogram discriminator. N and M denote the number of hidden layers in the diffusion discriminator and spectrogram discriminator, respectively. The mixing ratio λ is a hyperparameter that controls the sum of two expectation values.

$$\mathcal{L}_{fm} = \lambda \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t)} \left[\sum_{i=1}^{N} \| D_d^i\left(\mathbf{x}_{t-1},\ \mathbf{x}_t,\ t\right) - D_d^i\left(\mathbf{x}'_{t-1},\ \mathbf{x}_t,\ t\right) \|_1 \right] + (1-\lambda)\, \mathbb{E}_{\mathbf{x}_0 \sim p_{data}(\mathbf{x}_0), \mathbf{x}'_0 \sim p_\theta(\mathbf{x}_{0:T})} \left[\sum_{i=1}^{M} \| D_s^i\left(\mathbf{x}_0,\ s\right) - D_s^i\left(\mathbf{x}'_0,\ s\right) \|_1 \right], \tag{7}$$

In addition to the feature matching loss $\mathcal{L}_{fm}$, the variance adaptor is trained using the acoustic reconstruction loss $\mathcal{L}_{recon}$ following [20] to accurately predict key characteristics of the speech, such as duration, pitch, and energy. Finally, based on [20], we train the generator to minimize the adversarial loss:

$$\mathcal{L}_{adv} = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \left[(D_d\left(\mathbf{x}_{t-1},\ \mathbf{x}_t,\ t\right) - 1)^2 \right] \right] + \mathbb{E}_{\mathbf{x}'_0 \sim p_\theta(\mathbf{x}_{0:T})} \left[\left(D_s\left(\mathbf{x}'_0,\ s\right) - 1 \right)^2 \right]. \tag{8}$$

In total, the generator is trained by minimizing $\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}_{recon} + \lambda_{fm}\mathcal{L}_{fm}$, where λ_{fm} is a dynamically scaled scalar computed as $\lambda_{fm} = \mathcal{L}_{recon}/\mathcal{L}_{fm}$ at each training step, following [32]. Therefore, at each step, the feature matching loss is adjusted according to the ratio of the reconstruction loss to the feature matching loss.

IV. MODEL TRAINING AND EVALUATION

To validate the proposed SpecDiff-GAN model, we conducted multi-speaker speech synthesis experiments. In these experiments, we compared the proposed model with state-of-the-art alternatives to assess its ability to accurately capture speaker characteristics, generate smooth and natural pronunciation, and produce speech that closely resembles human-like sounds. The mel-spectrograms generated by SpecDiff-GAN were compared not only with the ground-truth mel-spectrograms but also with those generated by FastSpeech 2 [14] and DiffGAN-TTS [20]. To ensure a fair evaluation, all mel-spectrograms were converted into audio waveforms (i.e., speech) using the HiFi-GAN vocoder [33].

A. DATASET

The multi-speaker speech dataset used in this experiment is the voice cloning toolkit (VCTK) corpus [34]. The VCTK corpus, an open dataset created by the University of Edinburgh in Scotland, includes approximately 44 hours of English speech from 110 English-speaking speakers. Each speaker read 400 selected sentences from newspapers. All voices in the dataset were recorded in a non-reverberant indoor studio using high-performance microphones. The original recordings were made at a sampling rate of 96 kHz and in 24-bit audio format. They were subsequently downsampled to a sampling rate of 48 kHz and provided in 16-bit audio format. For the experiment, out of the total 44000 voice samples available,

512 were allocated as the validation set, while the remainder were used for training. Prior to being used in the experiments, all the data was further downsampled to a sampling rate of 22.05 kHz.

B. MODEL SETUP AND TRAINING

The proposed SpecDiff-GAN model was implemented using the PyTorch framework. The *librosa* library was utilized for loading audio files, while the *audio* library was employed for extracting mel-spectrograms and energy from the audio. Additionally, the *parselmouth* library was used to extract the fundamental frequency F0. In the conversion of audio to mel-spectrograms, we set the number of mel channels to 80, the hop size to 256, the window size to 1024, and the frequency range from 0 to 8000 Hz. The Adam optimizer was chosen for training.

Each transformer encoder in the generator has a hidden neuron count of 256. With four encoders, this configuration forms a feature space of 1024 dimensions. The encoder uses a convolution kernel size of 9. The variance adaptor comprises a duration predictor, pitch predictor, and energy predictor, each consisting of two convolution blocks. The padding sizes for each predictor are set to (1, 1), (2, 2), and (3, 3), respectively, and the kernel sizes are (3, 3), (5, 5), and (5, 5).

In the spectrogram discriminator, a zero-padding height of size 1 and a zero-padding width of size 4 are applied before performing convolution with a 3×9 kernel. Only the second convolutional layer applies a horizontal stride of 2. The main hyperparameters used in SpecDiff-GAN are summarized in Table 1.

The workstation used for training has the Ubuntu 20.04 LTS operating system, and software dependencies are managed using Docker. Training of the proposed model was conducted on four Nvidia A100 GPUs, each with 80 GB of memory. We utilized publicly available codes for FastSpeech 2 and DiffGAN-TTS, accessible via [35] and [36], respectively. All mel-spectrograms were converted into speech using a pre-trained HiFi-GAN vocoder. Our implementation and the audio samples used for evaluation are available on GitHub [37].

C. PERFORMANCE METRICS

The performance evaluation metrics used are SSIM [24], MCD [25], F0-RMSE, PER, WER, and RTF. While SSIM is a metric used in computer vision to measure the similarity of images, in this paper, it is used to compare the similarity between the real mel-spectrogram and the generated mel-spectrogram by considering spectrograms as images.

MCD is a metric used to measure the difference between the mel-cepstral coefficients of real and generated speech, expressing it in [dB], which indicates the quality of the speech. MCD is given by:

$$\mathrm{MCD}\left(\mathbf{x},\ \mathbf{x}'\right) = \frac{10}{\ln(10)}\sqrt{2\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{x}'_i\|_2}, \qquad (9)$$

where $\mathbf{x}$ and $\mathbf{x}'$ are mel-cepstrum of original and synthetic speeches, respectively, and N is the total number of frames of speeches. A lower value indicates a higher similarity to the real speech.

TABLE 1. Considered Hyperparameters for SpecDiff-GAN

Layer	Hyperparameters	Values
Generator	Number of diffusion time step	4
	Number of transformer encoder	4
	Transformer encoder hidden size	256
	Number of attention head	2
	Feed forward network hidden size	256
	Feed forward network filter size	1024
	Feed forward network kernel size	9
	Variance adaptor constant padding	{1, 2, 2}
	Variance adaptor kernel sizes	{3, 5, 5}
	Number of residual blocks	20
	Diffusion decoder kernel size	{3, 1, 1}
Diffusion discriminator	Mixing ratio (α)	0.5
	Number of convolution block	6
	Kernel size	{3}
Spectrogram discriminator	Number of Conv2d (strided)	3
	Number of Conv2d	2
	Kernel sizes	$\{3 \times 3, 3 \times 9\}$
	Stride height	1
	Stride widths	{1, 2}
	Padding height	1
	Padding widths	{1, 4}

F0-RMSE measures the difference between the ground-truth values and the generated values of the fundamental frequency F0. It is used to assess the accuracy of pitch in speech. A lower value indicates a better match between the ground-truth speech and the generated speech in terms of pitch. It is defined as:

$$\text{F0 RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F_i - F'_i\right)^2}, \qquad (10)$$

where N is the total number of frames of speeches, F_i is the F0 of the ground-truth speech, and F'_i is the F0 of the generated speech. A lower value indicates better performance in terms of F0 similarity between the ground-truth and generated speech.

PER and WER, assessing phoneme and word recognition accuracy respectively, are evaluated using the same methods as HierSpeech [26]. RTF represents the ratio of the total processing time of the synthesized speech to the duration of the synthesized speech. If the RTF is less than 1.0, it indicates that the system can generate speech in real-time. A lower RTF value means that the system's speech generation speed is faster.

V. RESULTS

The test set comprises voice recordings from English-speaking individuals, with each person speaking one of 512 sentences. The duration of each voice sample ranges from

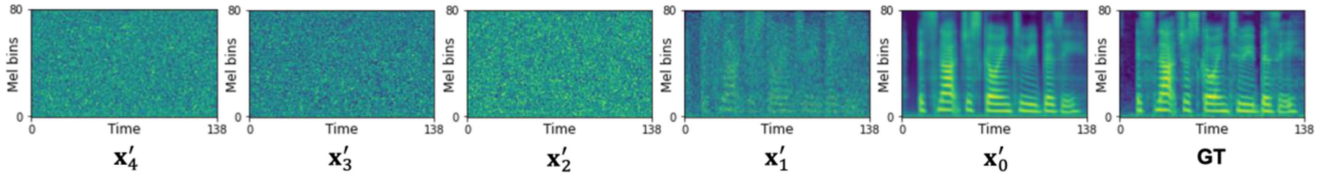

FIGURE 5. Visualization of the denoising process during inference of the proposed model, x'_0 is the generated mel-spectrogram, GT is the ground-truth mel-spectrogram.

TABLE 2. Objective Performance Evaluation of Speech Synthesis

Model	SSIM$_\uparrow$	MCD [dB]$_\downarrow$	F0-RMSE [Hz]$_\downarrow$	PER$_\downarrow$	WER$_\downarrow$	RTF$_\downarrow$
Mel-spectrogram (GT) + vocoder	-	**9.27** (±0.09)	-	**0.053** (±0.090)	**0.085** (±0.013)	-
FastSpeech 2 [14]	0.411 (±0.004)	11.96 (±0.08)	45.35 (±2.52)	0.075 (±0.012)	0.118 (±0.016)	**0.0035**
DiffGAN-TTS [20]	0.787 (±0.003)	9.50 (±0.09)	33.48 (±2.30)	0.105 (±0.014)	0.161 (±0.020)	0.0064
SpecDiff-GAN (ours)	**0.791** (±0.009)	9.35 (±0.07)	**33.19** (±2.23)	0.105 (±0.014)	0.158 (±0.020)	0.0063

All spectrograms are converted to audio using a pretrained HiFi-GAN Vocoder. Values in ( ) are 95% confidence intervals.
The best result are highlighted in bold.

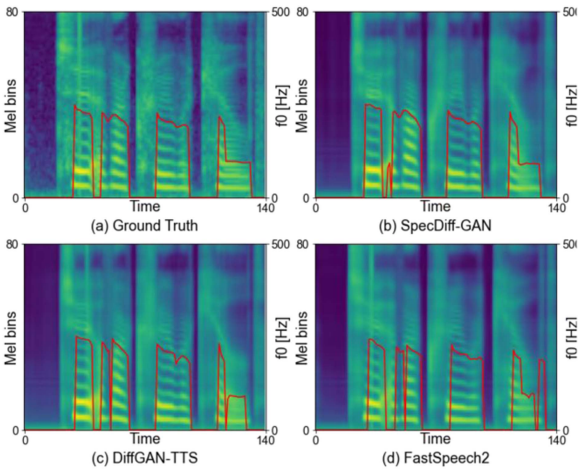

FIGURE 6. Comparison of mel-spectrogram with F0. (a) Ground Truth, (b) SpecDiff-GAN, (c) DiffGAN-TTS, and (d) FastSpeech 2.

TABLE 3. Subjective Performance Evaluation of Speech Synthesis

Model	SMOS$_\uparrow$	CMOS$_\uparrow$
Ground Truth (GT)	4.44 (±0.07)	-
Mel-spectrogram (GT) + vocoder	**4.36** (±0.05)	0.08 (±0.08)
FastSpeech 2 [14]	4.20 (±0.05)	-0.20 (±0.07)
DiffGAN-TTS [20]	4.31 (±0.06)	-0.06 (±0.07)
SpecDiff-GAN (ours)	4.35 (±0.06)	-

SMOS and CMOS Evaluation. The CMOS Score is the result of comparison with the proposed SpecDiff-GAN model. Values in ( ) are 95% confidence intervals.
The best result is highlighted in bold.

approximately 1 second to 9 seconds. For comparison purposes, DiffGAN-TTS and FastSpeech 2 were also trained under identical conditions. The proposed model underwent training for 300000 steps, taking approximately 49.2 hours in total. To observe the data generation process during the denoising phase in inference, we visualized the output at each timestep, as shown in Fig. 5. This visualization allows us to confirm that the generated output increasingly resembles the ground-truth mel-spectrogram as the denoising process progresses.

A. OBJECTIVE PERFORMANCE EVALUATION

Table 2 presents the evaluation of the performance of the models in generating mel-spectrograms. The time step for the diffusion model was set to 4 for all cases. As shown in Table 2, the proposed model demonstrates overall superior performance compared to the comparative models, especially in terms of MCD and F0-RMSE. These two metrics are indicators of comparing the characteristics and intonation of the speaker, making it evident that the spectrogram discriminator proposed in this paper effectively aids in generating mel-spectrograms that well learn the speaker and voice characteristics.

A high SSIM indicates that the mel-spectrograms generated by the proposed model closely resemble the ground-truth mel-spectrograms. Fig. 6 displays the mel-spectrograms and fundamental frequency (F0) of both real and generated speech. While it may be difficult to visually perceive differences between the generated mel-spectrograms, the F0 of the proposed

TABLE 4. Ablation Study Results

Model	SSIM↑	MCD [dB]↓	F0 RMSE [Hz]↓	PER↓	WER↓
Baseline: SpecDiff-GAN (ours)	**0.791** (±0.009)	**9.35** (±0.07)	**33.19** (±2.23)	0.105 (±0.014)	0.158 (±0.020)
without spectrogram discriminator with speaker embedding	0.779 (±0.009)	11.50 (±0.07)	40.58 (±2.24)	0.106 (±0.015)	0.165 (±0.021)
without spectrogram discriminator without speaker embedding	0.789 (±0.009)	11.63 (±0.07)	40.03 (±2.14)	0.111 (±0.016)	0.173 (±0.024)
DiffGAN-TTS [20] (revisited)	0.787 (±0.003)	9.50 (±0.09)	33.48 (±2.30)	0.105 (±0.014)	0.161 (±0.020)
FastSpeech 2 [14] (revisited)	0.411 (±0.004)	11.96 (±0.08)	45.35 (±2.52)	**0.075** (±0.012)	**0.118** (±0.016)

Comparison of the effects of different component in SpecDiff-GAN discriminator. Values in ( ) are 95% confidence intervals.
The best result are highlighted in bold.

model is observed to closely approximate the ground truth compared to the F0 of the comparative models.

In terms of PER and WER, the proposed model either shows slightly better or similar performance compared to DiffGAN-TTS [20], but these metrics are lower than those for FastSpeech 2 [14]. This discrepancy could be attributed to the use of separate discriminators for learning vocal features in DiffGAN-TTS, which more accurately replicate the pronunciation of the original speech. However, the limited number of timesteps in the denoising process might lead to some pronunciations being missed or inaccurately represented. In the case of the RTF metric, FastSpeech 2 achieved the highest measurement. The RTF of the proposed model, measured at 0.0063, indicates that it can generate speech much faster than human speech production speeds. This confirms that the proposed model is capable of maintaining a sufficiently fast generation speed while producing high-quality speech, making it well-suited for real-time applications.

B. SUBJECTIVE PERFORMANCE EVALUATION

Since all the mel-spectrograms were converted into speech using the same HiFi-GAN vocoder, this approach allows for a fair assessment of the quality of mel-spectrograms generated by the comparative models. Table 3 presents the evaluation results for the SMOS and CMOS, comparing the proposed model with the comparison models. We have invited 26 listeners to participate as evaluators for CMOS and SMOS. Among the evaluators, 17 are fluent English speakers, and the remaining participants are second-language users. SMOS is a method used to evaluate the quality of generated speech by presenting it to evaluators, who then rate its quality on a scale from 1 (worst) to 5 (best). A score of 1 indicates that the generated speech is severely degraded and almost unintelligible, while a score of 5 signifies speech that is noise-free, free from awkwardness, and has accurate pronunciation.

In the CMOS, with the proposed model serving as the reference, up to +3 points are assigned if the comparative model better represents the characteristics of the actual speech, and up to −3 points are awarded if it falls short in representing those characteristics. A score of 0 is given if there is no discernible difference between the proposed model and the reference speech. The original speech (i.e., ground truth) and the proposed model, which serves as the baseline, are not subject to CMOS evaluation. The evaluation was conducted using a diverse set of 30 speakers' voices, comprising a total of 150 audio clips provided to the evaluators. These audio clips consisted of unseen data, not used during the model training.

The SMOS scores reveal that the actual recorded voice obtained the highest rating, followed by the reconstructed speech using the mel-spectrogram from the original speech processed through a vocoder. Among the models generating speech from text, the proposed model outperformed FastSpeech 2 and DiffGAN-TTS, achieving higher scores. Therefore, it is evident that the proposed model generates more natural-sounding speech from text compared to the comparative models. This outcome demonstrates that the spectrogram discriminator in the proposed model positively impacts speech quality and the representation of speaker characteristics. The audio clips used for the subjective evaluation are available at [37].

C. ABLATION STUDY

Table 4 presents the results of an ablation study conducted to examine the impact of the spectrogram discriminator on the model's performance. Two experiments were conducted: the first involved completely removing the spectrogram discriminator, while the second entailed removing the spectrogram discriminator but connecting the speaker embedding, which was previously linked to the spectrogram discriminator, directly to the diffusion discriminator.

The experimental results confirmed that both models without the spectrogram discriminator exhibited a decrease in performance. Particularly, the model with the speaker embedding connected to the diffusion discriminator experienced even greater performance degradation. This observation suggests that learning both speaker information and multimodal distributions for the reverse process in a single discriminator adversely affects the model's performance. These results confirm that the presence of the spectrogram discriminator contributes to the improvement of the model's performance.

VI. CONCLUSION

In this paper, we proposed the SpecDiff-GAN model, a novel approach that combines a diffusion model with a GAN to enhance the quality of generated speech while maintaining efficient generation speeds. The proposed model features a dual discriminator structure, consisting of a diffusion discriminator and a spectrogram discriminator. The diffusion discriminator is responsible for modeling the multimodal distribution of the reverse process, thereby expediting the generation process of the diffusion model. In contrast, the spectrogram discriminator is specifically designed to discern between the original and generated mel-spectrograms. This enables the generator to produce high-quality speech that closely replicates the original. To evaluate the model's performance, we utilized five objective evaluation metrics alongside two subjective evaluation metrics. In the comparative analysis with FastSpeech 2 and DiffGAN-TTS, the proposed SpecDiff-GAN model demonstrated performance that was similar to or surpassed these models. Furthermore, the results of our ablation study confirmed the efficacy of the proposed techniques in enhancing the model's performance. The implementation of the dual discriminator structure has shown promising results in terms of performance improvement. Continuing research is underway to further refine the discriminator's performance.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[2] I. Goodfellow et al., "Generative Adversarial Networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[3] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1530–1538.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using non-equilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2000, pp. 1315–1318, doi: 10.1109/ICASSP.2000.861820.

[8] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–44010.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1983, pp. 804–807, doi: 10.1109/ICASSP.1983.1172092.

[12] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.

[13] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Workshop Speech Synth. Workshop*, 2016, p. 125.

[14] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–15.

[15] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6588–6592, doi: 10.1109/ICASSP39728.2021.9413889.

[16] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–20.

[17] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–28.

[18] 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.

[19] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 8067–8077.

[20] S. Liu, D. Su, and D. Yu, "DiffGAN-TTS: High-fidelity and efficient text-to-speech with denoising diffusion GANs," in *Proc. Int. Conf. Mach. Learn. Workshop Mach. Learn. Audio Synth.*, 2022.

[21] R. Huang, Y. Ren, Z. Jiang, C. Cui, J. Liu, and Z. Zhao, "FastDiff 2: Revisiting and incorporating GANs and diffusion models in high-fidelity speech synthesis," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 6994–7009, doi: 10.18653/v1/2023.findings-acl.437. [Online]. Available: https://aclanthology.org/2023.findings-acl.437

[22] Y. Deng, N. Wu, C. Qiu, Y. Luo, and Y. Chen, "MixGAN-TTS: Efficient and stable speech synthesis based on diffusion model," *IEEE Access*, vol. 11, pp. 57674–57682, 2023, doi: 10.1109/ACCESS.2023.3283772.

[23] Y. Ren, J. Liu, and Z. Zhao, "PortaSpeech: Portable and high-quality generative text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 13963–13974.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[25] R. Kubichek, "Mel-Cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128, doi: 10.1109/PACRIM.1993.407206.

[26] S. H. Lee, S. B. Kim, J. H. Lee, E. Song, M. J. Hwang, and S. W. Lee, "HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16624–16636.

[27] C. Li et al., "Deep Speaker: An end-to-end neural speaker embedding system," 2017, *arXiv:1705.02304*.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2226–2234.

[30] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proc. Interspeech*, 2021, pp. 2207–2211.

[31] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2813–2821, doi: 10.1109/ICCV.2017.304.

[32] J. Yang, J.-S. Bae, T. Bak, Y.-I. Kim, and H.-Y. Cho, "GANSpeech: Adversarial training for high-fidelity multi-speaker speech synthesis," in *Proc. Interspeech*, 2021, pp. 2202–2206, doi: 10.21437/Interspeech.2021-971.

[33] K. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.

[34] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)*. Chicago, IL, USA: Univ. Edinburgh, The Centre Speech Technol. Res., 2019.

[35] C. M. Chien, J. H. Lin, C. Y. Huang, P. C. Hsu, and H. Y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 8588–8592, doi: 10.1109/ICASSP39728.2021.9413880.

[36] K. Lee, "DiffGAN-TTS," Feb. 2022, doi: 10.5281/zenodo.6203248.

[37] M. Ko, "SpecDiff-GAN Implementation and audio samples," 2023. [Online]. Available: https://github.com/KoMyeongJin/SpecDiff-GAN

MYEONGJIN KO received the A.S. degree in computer and mobile convergence engineering from the Gyeonggi University of Science and Technology, Siheung-si, South Korea, in 2021, the B.S. degree in computer engineering from The Korean Academic Credit Bank System, Seoul, South Korea, in 2021, and the M.S. degree in robotics from Kwangwoon University, Seoul, in 2023.

His research interests include communication networks, software development, natural language processing, and speech recognition and synthesis.

EUIYEON KIM received the A.S. degree in computer and mobile convergence engineering from the Gyeonggi University of Science and Technology, Siheung-si, South Korea, in 2023, and the B.S. degree in computer engineering from The Korean Academic Credit Bank System, Seoul, South Korea, in 2023. He is currently currently working toward the M.S. degree in robotics with Kwangwoon University, Seoul.

His research interests include speech recognition and synthesis, audio source separation and enhancement, and machine learning model lightweighting.

YONG-HOON CHOI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1995, 1997, and 2001, respectively. From 2001 to 2002, he was a Research Associate with the Institute for Systems Research, University of Maryland, College Park, MD, USA. From 2002 to 2005, he was a Chief Research Engineer with LG Electronics. Since 2005, he has been a Faculty Member with Kwangwoon University, Seoul. From 2023 to 2024, he was the President of the Information Networking Society of the Korean Institute of Information Scientists and Engineers.

His research interests include communication networks, machine learning, speech synthesis, and AI-based FinTech. He was the recipient of the 28th Choon-Gang Award in 2013 from Choon-Gang Memorial Association.