IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2023, accepted 5 June 2023, date of publication 7 June 2023, date of current version 14 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283772

RESEARCH ARTICLE

MixGAN-TTS: Efficient and Stable Speech Synthesis Based on Diffusion Model

YAN DENG[1], **NING WU**[2], **CHENGJUN QIU**[3,4], **YANGYANG LUO**[1], **AND YAN CHEN**[1]

[1]School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[2]Key Laboratory of Beibu Gulf Offshore Engineering Equipment and Technology, Beibu Gulf University, Qinzhou 535011, China
[3]College of Mechanical Naval Architecture and Ocean Engineering, Beibu Gulf University, Qinzhou 535011, China
[4]Guangxi Key Laboratory of Ocean Engineering Equipment and Technology, Qinzhou 535011, China

Corresponding author: Ning Wu (n.wu@bbgu.edu.cn)

This work was supported in part by the China State Shipbuilding Corporation (CSSC) Guangxi Shipbuilding and Offshore Engineering Technology Collaboration Project under Grant ZCGXJSB20226300222-06, and in part by the 100 Scholar Plan of the Guangxi Zhuang Autonomous Region of China under Grant 2018.

ABSTRACT This paper describes MixGAN-TTS, an efficient and stable non-autoregressive speech synthesis based on diffusion model. The MixGAN-TTS uses a linguistic encoder based on soft phoneme-level alignment and hard word-level alignment approach which explicitly extracts word-level semantic information, and introduces pitch and energy predictors to optimally predict the rhythmic information of the audio. Specifically, we use the GAN to replace the Gaussian function to model the denoising distribution, aiming to enlarge the denoising steps size and reduce the number of denoising steps to accelerate the sampling speed of diffusion model. Diffusion model using GAN can significantly reduce the denoising steps, and to some extent solve the problem of not being able to apply in real-time. The mel-spectrogram is converted into the final audio by the HiFi-GAN vocoder. Experimental results show that the MixGAN-TTS outperforms the other models compared in terms of audio quality and mel-spectrogram modeling capability for 4 denoising steps. The ablation studies demonstrate that the structure of MixGAN-TTS is effective.

INDEX TERMS Speech synthesis, diffusion model, mixture attention mechanism, deep learning.

I. INTRODUCTION

As one of the core technologies of intelligent human-computer interaction, speech synthesis technology has been widely used in intelligent question and answer, intelligent navigation and so on. Speech synthesis, also known as text-to-speech (TTS), is a multimodal generation task that converts text to speech. The traditional TTS model consists of three key elements: text analysis frontend, acoustic model and neural vocoder [1], [2]. The text frontend normalizes the input text and converts the input text into linguistic representation features. The acoustic model converts the linguistic representation features into time-domain spectrogram acoustic features (e.g., mel-spectrogram). Finally, the time domain spectrogram acoustic features are converted into time domain waveform features by a neural vocoder. A large amount of research has been focused on vocoders in recent years [3], [4]. Common vocoders include WaveNet [5], HiFi-GAN [6], WaveGlow [7], and so on, have been widely used to generate human-like speech.

The autoregressive TTS model uses frame-by-frame prediction method and has demonstrated the ability to generate high quality audio. However, the autoregressive approach leads to slow training and inference, and has some robustness problems such as word skipping and repeating [8], [9], [10]. To solve these problems, non-autoregressive models have been proposed one after another. The FastSpeech model is one of the most successful non-autoregressive TTS models, which uses an encoder-decoder framework for processing the input phoneme sequences. FastSpeech introduces a duration predictor to obtain the duration of the training phonemes and obtains the alignment information between phoneme sequences and mel-spectrogram with the help of knowledge distillation [11]. The FastSpeech2 model introduces additional pitch and energy variance information and obtains alignment information between phoneme sequences and

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

mel-spectrogram with the help of the Montreal forced alignment (MFA), simplifying the training process and achieving a large improvement in audio quality [12]. The Glow-TTS model uses normalized flow and dynamic programming methods to directly search for the most likely monotonic alignment information between the phoneme sequences and mel-spectrogram [13]. The PortaSpeech model uses soft-level phoneme, hard-level word alignment, Variational Autoencoder (VAE), and normalized flow method that enables high quality audio and expressive features [14].

Another generative model using denoising diffusion probabilistic models (DDPMs) has obtained satisfactory results [15], [16], [17], [18], [19]. Denoising diffusion probability models, or diffusion models, have proven to have powerful modeling capabilities in areas such as music synthesis and image synthesis [20], [21]. The traditional diffusion model is divided into a diffusion process and a denoising process. The diffusion process adds small random noise to the data through a parameter-free T-step Markov chain. The denoising process gradually removes the added noise by a parameterized T-step Markov chain. Although the diffusion model exhibits strong modeling capabilities, it suffers from slow sampling speed and requires a large enough number of denoising steps, making it difficult to use for real-time applications. Traditional diffusion models use a Gaussian distribution to approximate the true denoising distribution in the denoising process and assume a small value of predefined variance in the Gaussian distribution. Therefore, when the real data is complex, it is not possible to model the noise information by a simple Gaussian distribution, which will affect the quality of the synthesized audio and the speed of inference [21]. To enlarge the denoising steps size, the diffusion model uses conditional generative adversarial networks (GAN) as a non-Gaussian distribution function to model the denoising distribution [15], [21].

To optimize the alignment information between phoneme sequences and mel-spectrogram and to improve the quality of model synthesized audio. Inspired in part by the TTS models [14], [15], this paper proposes a non-autoregressive model MixGAN-TTS. MixGAN-TTS model, which combines some structures of PortaSpeech and DiffGAN-TTS models and improves them accordingly. The MixGAN-TTS addresses the hard-level phoneme alignment problem of the FastSpeech2 by introducing a linguistic encoder with pitch and energy information, which uses soft-level phoneme, hard-level word alignment method (mixture alignment mechanism) can effectively solve the hard-level phoneme alignment problem caused by the MFA. Moreover, for the powerful modeling capability of diffusion models and the advantage of GAN that can cope with complex data distributions, we introduce the use of GAN as diffusion decoder for denoising distribution modeling methods, aiming to solve the problem that diffusion model exists which requires a large enough number of denoising steps. We evaluate the MixGAN-TTS on the AISHELL3 dataset [22], and the experimental results show that MixGAN-TTS achieves notable results in terms of synthesized audio quality, predicted mel-spectrogram, and attention alignment.

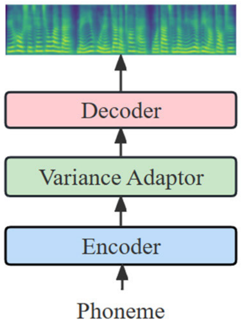


FIGURE 1. The overall architecture for FastSpeech2.

II. BACKGROUND

In this section, we first introduce the non-autoregressive model FastSpeech2, and then we introduce the diffusion model.

A. FASTSPEECH2

The FastSpeech2 is based on FastSpeech, and the alignment information between phoneme sequences and mel-spectrogram is obtained with the help of the MFA to solve the problem of increased training cost of the FastSpeech using the teacher-student model. As shown in Figure 1, FastSpeech2 introduces the Variance Adaptor which consists of a duration predictor, a pitch predictor and an energy predictor. The Variance Adaptor extracts variance information such as duration, pitch and energy from real audio in the training, and provides rich variance information as input conditions to improve the quality of synthesized audio in the inference. FastSpeech2 uses the feed-forward transformer (FFT) as the basic structure of the encoder and decoder, which consists of a self-attention mechanism and 1D convolution network. The audio quality has also achieved good results.

B. DIFFUSION MODEL

The diffusion model is divided into a diffusion process and a denoising process. Diffusion process, also known as the forward process, refers to the complete collapse of the data by gradually adding noise to the data until after T steps. Diffusion process is accompanied by predefined variance information β_t, and variance information $\beta_{1:T}$ employed at each step is independently distributed. Diffusion process is shown in equations (1) and (2), and the diffusion process gradually adds noise information to x_0, and iterates $x_{1:T}$ in turn to obtain the collapse data x_T:

$$q\left(x_{1:T} \mid x_0\right) = \prod\nolimits_{t\geq 1} q\left(x_t \mid x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) = N(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}) \tag{2}$$

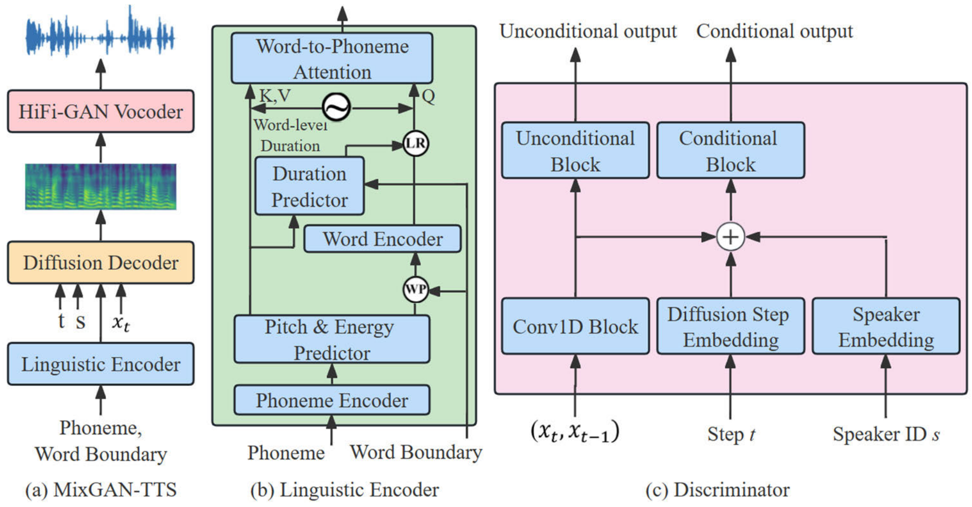


FIGURE 2. The overall architecture for MixGAN-TTS.

Denoising process, also known as the inverse process, gradually removes the noise information from the collapsed data by defining a denoising function. As shown by equations (3) and (4):

$$p_\theta\left(x_{0:T}\right) = p\left(x_T\right) \prod\nolimits_{t\geq 1} p_\theta\left(x_{t-1} \mid x_t\right) \tag{3}$$

$$p_\theta\left(x_{t-1} \mid x_t\right) = N\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sigma_t^2 \mathbf{I}\right) \tag{4}$$

The $p_\theta\left(x_{t-1} \mid x_t\right)$ denoising distribution is usually modeled using a Gaussian distribution, $\mu_\theta\left(x_t, t\right)$ and σ_t^2 denote the mean and variance of the denoising function, and θ denotes the parameter of the denoising function. Denoising process gradually and iteratively denoises $x_{T-1:0}$ from the Gaussian noise x_T to obtain the final generated data x_0. The training objective of the diffusion model is to maximize the $p_\theta(x_0)$ likelihood, which is achieved by maximizing the evidence lower bound (ELBO $\leq \log p_\theta(x_0)$). The diffusion model is optimized by ELBO to force the parameterized denoising model $p_\theta\left(x_{t-1} \mid x_t\right)$ to match the true denoising distribution $q\left(x_{t-1} \mid x_t\right)$. The ELBO is shown in equations (5) and (6):

$$\text{ELB0} = \sum\nolimits_{t\geq 1} E_{q(x_t)} D_{KL} + C \tag{5}$$

$$D_{KL} = D_{KL}\left(q\left(x_{t-1} \mid x_t\right) || p_\theta\left(x_{t-1} \mid x_t\right)\right) \tag{6}$$

D_{KL} denotes the relative entropy, also known as the Kullback-Leibler (KL) scatter. C represents the constant term that does not depend on the θ parameter.

III. MIXGAN-TTS

This paper proposes the MixGAN-TTS model, which aims to improve the alignment information between the phoneme sequences and mel-spectrogram and audio quality. In this section, we first describe the motivation of MixGAN-TTS, model composition structure, and then we describe the training losses of the MixGAN-TTS.

A. MOTIVATION

FastSpeech2 uses the MFA to obtain the alignment mechanism between the phoneme sequences and mel-spectrogram, ignoring the fact that phonemes have no obvious boundaries in the mel-spectrogram distribution, which will lead to a boundary blurring problem between different phonemes in the alignment process. To address the phoneme-level hard alignment problem of the FastSpeech2, a mixture alignment mechanism based linguistic encoder is introduced in PortaSpeech. PortaSpeech has demonstrated satisfactory model performance. In this work, we aim to further optimize the performance of the PortaSpeech model. We investigate pitch and energy variance information in the linguistic encoder structure. Experiments show that the introduced pitch and energy predictors can optimize the rhythmic information of the synthesized audio and obtain better audio quality.

Currently, diffusion models using the GAN to model denoising distributions have achieved significant results. Diffusion model based on GAN training discriminator to force the predicted denoising model distribution $p_\theta\left(x_{t-1} \mid x_t\right)$ to approximate the true denoising distribution $q\left(x_{t-1} \mid x_t\right)$, which can effectively extract the rich information of the true sample. Therefore, we design the MixGAN-TTS based on the diffusion model. The MixGAN-TTS architecture is shown in Figure 2(a), which uses a linguistic encoder based on mixture alignment mechanism, a diffusion decoder and a discriminator as the basic structure. In the next few subsections, we will introduce the structure of the MixGAN-TTS and describe the training losses of the MixGAN-TTS.

B. LINGUISTIC ENCODER

The linguistic encoder structure is shown in Figure 2(b), "LR" denotes the length regulator proposed in FastSpeech, "WP" denotes the word-level pooling operation proposed in PortaSpeech and Sinusoidal-like symbol denotes the positional encoding [23]. The linguistic encoder contains a phoneme encoder, a pitch predictor, an energy predictor, a word encoder, a duration predictor, and a word-phoneme attention module. Both the phoneme encoder and word encoder have forward feedback blocks as the basic structure. We first input Chinese characters and get the corresponding phoneme sequences (e.g., n i3 h ao3 sh iii4 j ie4). And then we normalize the phoneme sequences to identify aspects of Chinese characters such as polysyllabic characters, numbers and proper nouns. The linguistic encoder receives phoneme sequences with word boundaries (e.g., n i3 h ao3 | sh iii4 j ie4, "|" denotes word boundaries in the phoneme sequences) and encodes them as phoneme hidden states. The pitch and energy predictors extract extra variance information from the real audio for training, and then add the variance information to the phoneme hidden states in the inference stage. The phoneme hidden states are applied by the word-level pooling to obtain the input representation of the word encoder, which averages the phoneme hidden states inside each word according to the word boundary. The word encoder encodes the input representation as the word-level hidden states and expand them to match the length of the target mel-spectrogram using a length regulator with word-level duration. Finally, the word-phoneme attention module takes the word-level hidden states as the query Q and the phoneme hidden states as the key K and the value V. The phoneme hidden states and word-level hidden states are then encoded with word-level relative position and fed them into the word-phoneme attention module.

C. DIFFUSION DECODER AND DISCRIMINATOR

The true denoising distribution $q(x_{t-1} \mid x_t)$ is usually unknown, so it is difficult to calculate the KL scatter directly as shown in equation (5) [16]. Diffusion model uses a Gaussian function distribution to model the noise information in the diffusion stage. To calculate the KL scatter, denoising process tends to take smaller predefined variance information β_t and a large enough number of denoising steps T, which forces the denoising process to use the same Gaussian function modeling approach as the diffusion process [16]. We introduce the same diffusion decoder [15] and use conditional GAN to model the denoising distribution by increasing the denoising steps size and decreasing the number of denoising steps. High quality audio generation can be obtained with only a few denoising steps. The ablation study shows that the best results are achieved with the MixGAN-TTS when T is equal to 4. The structure of the diffusion decoder is shown in Figure 3. The basic structure uses non-causal residual blocks, and the output of the residual blocks is used as additional input conditions for the subsequent structures through residual connections. Each residual block has a hidden state dimension of 256 dimensions and total of 20 blocks. The diffusion decoder takes the output of the linguistic encoder, diffusion step coding, noise mel-spectrogram x_T and speaker embedding as input conditions. The output of the linguistic encoder is passed into the residual block through a 1D convolution network, and x_T is diffusion step coded through the residual connection by a 1D convolution network and ReLU activation function. MixGAN-TTS uses relative position encoding on the diffusion steps t, which is accessed into the diffusion decoder via the full connection layer (FC) and the swish activation function [24]. The speaker embedding as an independent input condition aims to generate multi-speaker style audio, which is passed into the residual block via a 1D convolution network. The residual block introduces a gating mechanism [5] making full use of the Tanh and Sigmoid activation function. Finally, the final diffusion decoder output is obtained by sequentially accessing the output of the residual blocks through the skip connection and alternately using the 1D convolution network and ReLU activation function.

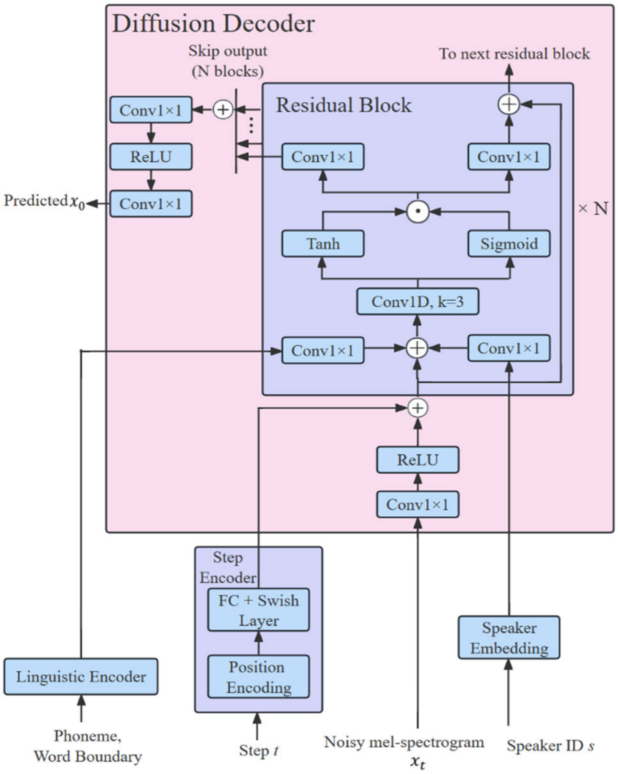


FIGURE 3. **The architecture for diffusion decoder.**

The discriminator structure is shown in Figure 2(c), with noise information x_t, predicted spectrogram x_{t-1}, diffusion step embedding t and speaker embedding s as input conditions, aiming to calculate the convergence degree D_{adv} between the true denoising distribution $q(x_{t-1} \mid x_t)$ and the denoising model distribution $p_\theta(x_{t-1} \mid x_t)$ during each denoising step. We use the least-squares GAN [25] to train the discriminator. The discriminator structure is modeled and represented by $D_\varphi(x_{t-1}, x_t, t, s)$ with learnable parameters φ. The discriminator uses joint conditional and unconditional loss (JCU) [26], which combines conditional and unconditional adversarial losses to further improve the accuracy of the mel-spectrogram and speech waveform mapping.

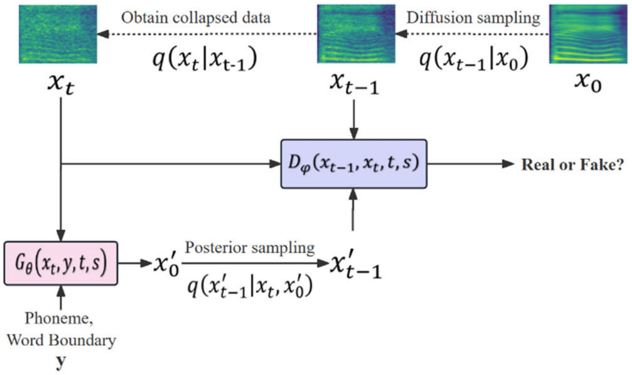


FIGURE 4. Training process of MixGAN-TTS.

D. TRAINING AND INFERENCE

The MixGAN-TTS training process as show in Figure 4, MixGAN-TTS samples the real mel-spectrogram x_0 to obtain the noise spectrogram x_t. The linguistic encoder and the diffusion decoder are modeled by the generator $G_\theta\left(x_t, y, t, s\right)$ parameterized with θ. The generator takes x_t noise sample, phoneme sequences y, diffusion step sequences t and speaker embedding s as input conditions and obtains the predicted mel-spectrogram map x'_0 through a T step denoising process. In the training stage, x'_0 is obtained as the predicted mel-spectrogram x'_{t-1} by the posterior distribution $q\left(x'_{t-1} \mid x_t, x'_0\right)$, and is passed into the JCU discriminator together with the noise spectrum x_t to calculate the convergence degree of the denoising process D_{adv}. In the inference stage, the sample x'_0 is generated by the generator, and the final predicted mel-spectrogram is obtained by the process of diffusion sampling and denoising. Finally, we use the HiFi-GAN vocoder to convert the mel-spectrogram into speech waveform.

E. TRAINING LOSS

The MixGAN-TTS training loss consists of generator loss and discriminator loss. We use the feature matching loss L_{fm} [27], acoustic reconstruction loss L_{recon} and denoising convergence loss L_{adv} to train the generator together. L_{fm} learns the similarity measure to distinguish the real data from the generated data in the discriminator. As shown in equation (7), L_{fm} is computed by summing the l_1 distance between the real and generated samples in the discriminator:

$$L_{fm} = E_{q(x_t)}\left[\sum_{i=1}^{N} D_\varphi^i\left(x_{t-1}, x_t, t, s\right) - D_\varphi^i\left(x'_{t-1}, x_t, t, s\right) \|_1\right] \tag{7}$$

N denotes the number of hidden layers of the discriminator. L_{recon} calculates the basis reconstruction loss, and L_{adv} calculates the convergence degrees between the true denoising distribution $q\left(x_{t-1} \mid x_t\right)$ and the denoising model distribution $p_\theta\left(x_{t-1} \mid x_t\right)$, as shown in equations (8) and (9):

$$\begin{aligned} L_{recon} = {} & L_{mel} + \lambda_d L_{duration} + \lambda_p L_{pitch} \\ & + \lambda_e L_{energy} + L_{helper} \end{aligned} \tag{8}$$

$$L_{adv} = \sum\nolimits_{t\geq 1} E_{q(x_t)} E_{p_\theta\left(x_{t-1}, x_t\right)}\left[\left(D_\varphi\left(x_{t-1}, x_t, t, s\right) - 1\right)^2\right] \tag{9}$$

where λ_d, λ_p and λ_e denote loss weights set to 0.1. L_{mel} uses MAE loss, $L_{duration}$, L_{pitch} and L_{energy} use MSE loss, and L_{helper} uses Guided Attention Loss [28]. The generator is trained by minimizing L_G:

$$L_G = L_{fm} + L_{recon} + L_{adv} \tag{10}$$

The discriminator is trained by minimizing the L_D loss:

$$\begin{aligned} L_D = {} & \sum_{t\geq 1} E_{q(x_t)q\left(x_{t-1} \mid x_t\right)}\left[\left(D_\varphi\left(x_{t-1}, x_t, t, s\right) - 1\right)^2\right] \\ & + E_{p_\theta\left(x_{t-1} \mid x_t\right)}\left[D_\varphi\left(x_{t-1}, x_t, t, s\right)^2\right] \end{aligned} \tag{11}$$

IV. EXPERIMENTS AND RESULTS

To evaluate the audio modeling performance and audio quality of the MixGAN-TTS, we design the comparison experiments between FastSpeech2, PortaSpeech, DiffGAN-TTS ($T = 4$) and DiffGAN-TTS (two-stage) and MixGAN-TTS. In this section, we first introduce the datasets and the model configuration, and then we describe the evaluation methods and experimental results. Finally, the modules added to MixGAN-TTS are studied for ablation to verify the effectiveness of each structure.

A. DATASETS

We evaluate the performance of the MixGAN-TTS on the AISHELL3 dataset, which is recorded by 218 Mandarin speakers and contains 88,035 Chinese audio clips and corresponding text transcripts. We divide the dataset into three subsets: 87011 samples for training, 512 samples for validation, and 512 samples for testing. We randomly select 160 samples from the test dataset for objective evaluation and 20 samples from the test dataset for subjective evaluation. We use the pypinyin library to convert text sequences to phoneme sequences, and convert the original waveform to mel-spectrogram at a sampling rate of 22050Hz with 16bit sampling bits. The mel-spectrogram has a window size of 1024 and a hop count of 256. All sentences in the dataset are pre-processed to remove gaps before and after the audio, as well as to normalize the text.

B. MODEL CONFIGURATION

We train the MixGAN-TTS model on an NVIDIA 3060 GPU, setting the processing batch to 8 and using the Adam optimizer with parameters set to $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\epsilon = 10^{-9}$ both the generator and discriminator followed by [14] and [15]. We set the learning rate of gradual decay, and the initial values of the learning rate of the generator and discriminator are set to 10^{-3} and 2×10^{-3} respectively. The CUDA version

TABLE 1. Model experiment evaluation and model efficiency results. Higher SSIM values are better, while lower values of MCD and F_0 RMSE are better.

Model	SSIM	MCD	F_0 RMSE	Params	RTF	MOS	CMOS
Ground Truth						4.17±0.07	
FastSpeech2	0.494	17.348	0.713	30.87M	0.2318	3.76±0.08	0.000
PortaSpeech	0.510	17.146	**0.706**	24.26M	0.2368	**3.91±0.07**	0.149
DiffGAN-TTS(T=4)	0.506	17.248	0.739	29.04M	0.2281	3.86±0.08	0.134
DiffGAN-TTS(two-stage)	0.508	17.219	0.802	40.25M	0.2292	3.85±0.08	0.137
MixGAN-TTS	**0.511**	**17.127**	0.764	25.31M	0.2263	3.89±0.09	0.144

for all the experiments is 11.6, and the model programming utilizes python 3.8 with pytorch version 1.8.0+cu111. MixGAN-TTS is trained for at least 900k steps until losses converge. We use the HiFi-GAN vocoder publicly trained in github to transform the mel-spectrogram into audio samples. We randomly select some of the test dataset for mean opinion score (MOS) [29] and comparative mean opinion score (CMOS) [30] tests. We keep the text content consistent across models, exclude other confounding factors, and check only the audio quality. Each audio sample are rated by a minimum of 10 testers.

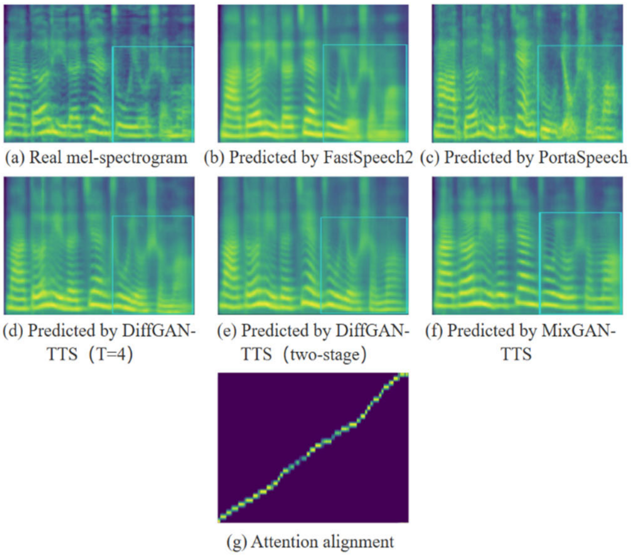
C. EVALUATE

To measure the quality and performance of the model synthesized audio, we use structural similarity index (SSIM) [31], mel-cepstral distortion (MCD) [32] and F_0 root mean squared error (F_0 RMSE) metrics for objective evaluation of the model, and MOS and CMOS metrics for subjective evaluation of the model. The higher the SSIM value, the closer the synthesized spectrogram is to the real mel-spectrogram, indicating that the synthesized audio is closer to the original audio to some extent. Dynamic time warping (DTW) [33] is used for MCD and F_0 RMSE calculations to align the generated audio with the real reference audio. In our work, the logarithmic method is used to calculate the F_0 RMSE value. The lower values of MCD and F_0 RMSE indicate better quality of the synthesized audio to some extent. As shown in Table 1, the MixGAN-TTS proposed achieves the best SSIM and MCD values. In terms of F_0 RMSE evaluation metrics, MixGAN-TTS is lower than the DiffGAN (two-stage) model and slightly higher than the FastSpeech2, PortaSpeech and DiffGAN-TTS ($T = 4$) models. The experimental results show that the MixGAN-TTS model achieves satisfactory performance in alignment information between the phoneme sequences and mel-spectrogram, and is able to achieve high quality audio, as we can also find in terms of subjective evaluation and audio samples. Params denotes the number of parameters of the model. From the results in Table 1, it can be seen that the MixGAN-TTS achieves significant results in optimizing the number of parameters, which is 25.31M, much less than the rest of the compared models. In addition, we use the real- time factor (RTF) as a measure of the model inference speed. RTF indicates the time required by the model to generate on second of audio. The smaller the RTF value, the faster the model synthesizes audio. We select 20 generated audio samples for RTF testing, with durations ranging from 3 to 6 seconds and the number of Chinese words ranging from 8 to 20 characters. RTF test results are shown in Table 1. All models are trained and inferred on an NVIDIA 3060 GPU.

To assess the quality of the synthesized audio, a sample of 20 sentences are randomly selected from the test dataset for MOS subjective evaluation with a set confidence interval of 95%. Each audio sample is evaluated by at least 10 testers. We conduct the test in a quiet classroom and assign a dozen testers to score the test. The process from configuration to testing took several hours. As a result, there are fluctuations in the number of people. We break up all samples involved in the test. No label is given as to which model the sample is generated from. Testers are asked to score each speech carefully, ranging from 1 to 5 with a 0.5-point increment, in terms of speech naturalness and accent performance. All testers are native Chinese and wear headphones for the test. As can be observed from the data in Table 1, MixGAN-TTS achieves an MOS value of 3.89, which is better than FastSpeech2, DiffGAN-TTS ($T = 4$) and DiffGAN-TTS (two-stage), and comparable to the audio quality of PortaSpeech, which indicates that the MixGAN-TTS is able to better learn the alignment information between the phoneme sequences and mel-spectrogram, and the synthesized audio quality is better. In addition, CMOS is used to compare the performance between the models. FastSpeech2 model is used as a baseline, and testers listen to the audio generated by each model separately for comparison. From the results in Table 1, it is observed that the audio quality of the MixGAN-TTS model is better than that of the FastSpeech2 and achieves audio quality comparable to that of the PortaSpeech, DiffGAN-TTS ($T = 4$), and DiffGAN (two-stage) models.

D. FEATURE PREDICTION

We analyze the mel-spectrogram and attention alignment mechanism. The mel-spectrogram predicted by different models are given in Figure 5. Combining the SSIM evaluation metrics as shown in Table 1 and the predicted mel-spectrogram, we can observe the MixGAN-TTS has an advantage in predicting the mel-spectrogram details in frequency bins between two adjacent harmonics, unvoiced frames and low-frequency parts, and the model generates better quality audio. Figure 5(g) gives the attention alignment convergence of the MixGAN-TTS during the training


(a) Real mel-spectrogram (b) Predicted by FastSpeech2 (c) Predicted by PortaSpeech

(d) Predicted by DiffGAN-TTS (T=4) (e) Predicted by DiffGAN-TTS (two-stage) (f) Predicted by MixGAN-TTS

(g) Attention alignment

FIGURE 5. Feature prediction.

TABLE 2. CMOS comparison for MixGAN-TTS.

Setting	CMOS
MixGAN-TTS	0.000
MixGAN-TTS - pitch	-0.231
MixGAN-TTS - energy	-0.176
MixGAN-TTS - pitch - energy	-0.316

process. The MixGAN-TTS achieves better results in the alignment information between the phoneme sequences mel-spectrogram, the brightness and lines of the points in the alignment map are clear, and the convergence map is clear and smooth, showing that the MixGAN-TTS has a notable ability to align the phoneme sequences with the mel-spectrogram map.

E. ABLATION STUDIES

To verify the effectiveness of the structure of the MixGAN-TTS, an ablation study of the pitch and energy variance information introduced in the linguistic encoder module and the also diffusion decoder module is conducted. Comparison results of the ablation study of the variance information introduced by the model are given in Table 2. The MixGAN-TTS leads to a degradation of audio quality after the removal of pitch and energy information. The CMOS value of the MixGAN-TTS with the pitch information removed is -0.231; with the energy information removed, the CMOS value is 0.176; and the CMOS value of the MixGAN-TTS model with both pitch and energy removed is -0.316. Experimental results show that the MixGAN-TTS can improve the audio quality by introducing pitch and energy variance information in the linguistic encoder, and the model can learn the pitch and energy variance information of real audio in the training stage, which provides rich variance information for the model in the inference stage.

TABLE 3. Ablation studies comparison for diffusion decoder.

Setting	MCD	F_0 RMSE	CMOS
MixGAN-TTS	17.127	0.764	0.000
MixGAN-TTS - diffusion	17.426	0.713	-0.231

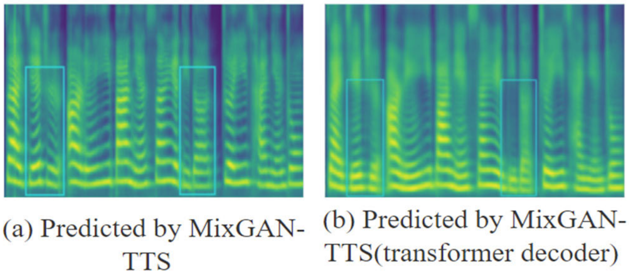

(a) Predicted by MixGAN-TTS (b) Predicted by MixGAN-TTS(transformer decoder)

FIGURE 6. The mel-spectrogram comparison for ablation studies of diffusion decoder.

TABLE 4. Ablation studies comparison for denoise step T.

Setting	MCD	F_0 RMSE	CMOS
MixGAN-TTS(T=1)	17.395	0.781	0.000
MixGAN-TTS(T=2)	17.321	0.769	0.144
MixGAN-TTS(T=4)	17.127	0.764	0.263

To verify the effectiveness of the diffusion decoder module, we use MCD, F_0 RMSE and CMOS metrics for the study. The comparison models are structured as a linguistic encoder with pitch and energy information and a transformer decoder in the FastSpeech2 model. As can be observed from the results in Table 3, the introduction of the diffusion decoder can effectively improve the quality and model performance of the synthesized audio, while the number of parameters is also effectively introduced, indicating that the diffusion decoder can take full advantage of the powerful modeling capabilities of adversarial generative networks to resolve complex data distributions and alignment information between phoneme sequences and mel-spectrogram.

In addition, we investigate the mel-spectrogram before and after the introduction of the diffusion decoder module. Figure 6(a) shows the mel-spectrogram predicted by the MixGAN-TTS, and Figure 6(b) shows the mel-spectrogram predicted by the MixGAN-TTS (Transformer decoder) model. It can be observed from the figure that the MixGAN-TTS with the introduction of the diffusion decoder predicts a richer internal detail and better speech quality of the mel-spectrogram.

We investigate the number of denoising steps T, as shown in Table 4, and the results show that the best MCD and F_0 RMSE results were obtained when $T = 4$, with 17.127 and 0.764, respectively, and the best audio quality. The MixGAN-TTS is gradually diffusion sampled from the initial data x_0 to get the collapse data x_4, and x_4 is passed into the generator as noise information for training, and the gradual denoising sampling $x'_{3:1}$ gets the final predicted mel-spectrogram x'_0.

V. DISCUSSION

In this paper, we propose the non-autoregressive speech synthesis MixGAN-TTS model, which combines some structures of PortaSpeech and DiffGAN-TTS models and improves them accordingly. The MixGAN-TTS solves the phoneme boundary ambiguity problem of hard-level phoneme alignment, introduces a mixture alignment mechanism based on linguistic encoder, and adds pitch and energy predictors to further handle the variance information of real audio. In addition, the original diffusion model uses a Gaussian function distribution to model the denoising distribution, which is limited by the small denoising steps size and the large number of denoising steps, resulting in poor real-time performance of the diffusion model. MixGAN-TTS uses the GAN to model the real denoising distribution, which can generate high-quality audio with a large number of denoising steps size and a small number of denoising steps. We use a discriminator to calculate the convergence of noise information between the sample distribution predicted by the generator and the real noise distribution, and synthesize the audio output using the HiFi-GAN vocoder.

We perform a subjective and objective evaluation of the MixGAN-TTS. MOS and CMOS metrics are used for the subjective side, and then SSIM, MCD and F_0 RMSE metrics are used for the objective side. MixGAN-TTS obtain the best SSIM and MCD values, but is slightly weaker in terms of F_0 RMSE. In addition, we conduct an ablation study on the structure of the MixGAN-TTS model to show the validity of each part of the model. Experimental results show that the MixGAN-TTS achieves the best results with 4 denoising steps, and the mel-spectrogram reconstruction and audio quality are both improved significantly. MixGAN-TTS model is not a true end-to-end model, it still needs to be converted to speech waveform with the help of vocoder. A full end-to-end model will be our future research plan.

REFERENCES

[1] W. Zhang, H. Yang, X. Bu, and L. Wang, ''Deep learning for mandarin-tibetan cross-lingual speech synthesis,'' *IEEE Access*, vol. 7, pp. 167884–167894, 2019.

[2] D. Panagiotopoulos, C. Orovas, and D. Syndoukas, ''Neural network based autonomous control of a speech synthesis system,'' *Intell. Syst. Appl.*, vol. 14, May 2022, Art. no. 200077.

[3] M. Salah Al-Radhi, T. Gábor Csapó, C. Zainkó, and G. Németh, ''Continuous wavelet vocoder-based decomposition of parametric speech waveform synthesis,'' 2021, *arXiv:2106.06863*.

[4] M. S. Al-Radhi, T. G. Csapó, and G. Németh, ''Continuous vocoder applied in deep neural network based voice conversion,'' *Multimedia Tools Appl.*, vol. 78, no. 23, pp. 33549–33572, Dec. 2019.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, ''WaveNet: A generative model for raw audio,'' 2016, *arXiv:1609.03499*.

[6] J. Kong, J. Kim, and J. Bae, ''HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17022–17033.

[7] R. Prenger, R. Valle, and B. Catanzaro, ''WaveGlow: A flow-based generative network for speech synthesis,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, ''Tacotron: Towards end-to-end speech synthesis,'' 2017, *arXiv:1703.10135*.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, ''Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[10] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, ''Neural speech synthesis with transformer network,'' in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 6706–6713.

[11] Y. Ren, Y. Runa, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, ''FastSpeech: Fast, robust and controllable text to speech,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–22.

[12] 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,'' 2020, *arXiv:2006.04558*.

[13] 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.*, vol. 33, 2020, pp. 8067–8077.

[14] Y. Ren, J. Liu, and Z. Zhao, ''PortaSpeech: Portable and high-quality generative text-to-speech,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13963–13974.

[15] S. Liu, D. Su, and D. Yu, ''DiffGAN-TTS: High-fidelity and efficient text-to-speech with denoising diffusion GANs,'' 2022, *arXiv:2201.11972*.

[16] J. Ho, A. Jain, and P. Abbeel, ''Denoising diffusion probabilistic models,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[17] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, ''DiffSinger: Singing voice synthesis via shallow diffusion mechanism,'' in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 11020–11028.

[18] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, ''Wavegrad: Estimating gradients for waveform generation,'' 2020, *arXiv:2009.00713*.

[19] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, ''Grad-TTS: A diffusion probabilistic model for text-to-speech,'' in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8599–8608.

[20] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, ''Symbolic music generation with diffusion models,'' 2021, *arXiv:2103.16091*.

[21] Z. Xiao, K. Kreis, and A. Vahdat, ''Tackling the generative learning trilemma with denoising diffusion GANs,'' 2021, *arXiv:2112.07804*.

[22] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, ''AISHELL-3: A multi-speaker Mandarin TTS corpus and the baselines,'' 2020, *arXiv:2010.11567*.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[24] P. Ramachandran, B. Zoph, and Q. V. Le, ''Searching for activation functions,'' 2017, *arXiv:1710.05941*.

[25] 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. (ICCV)*, Oct. 2017, pp. 2813–2821.

[26] J. Yang, J.-S. Bae, T. Bak, Y. Kim, and H.-Y. Cho, ''GANSpeech: Adversarial training for high-fidelity multi-speaker speech synthesis,'' 2021, *arXiv:2106.15153*.

[27] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, ''Autoencoding beyond pixels using a learned similarity metric,'' in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[28] H. Tachibana, K. Uenoyama, and S. Aihara, ''Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4784–4788.

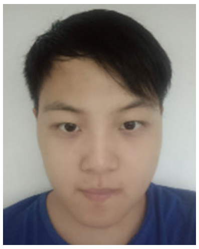
[29] M. Chu and H. Peng, ''Objective measure for estimating mean opinion score of synthesized speech,'' U.S. Patent 7 024 362, 2006.

[30] P. C. Loizou, ''Speech quality assessment,'' in *Multimedia Analysis, Processing and Communications*. Berlin, Germany: Springer, 2011, pp. 623–654.

[31] 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.

[32] 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.

[33] M. Müller, ''Dynamic time warping,'' in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.



YAN DENG received the B.E. degree from the Hunan Institute of Technology, in 2021. He is currently pursuing the master's degree with the School of Computer and Electronic Information, Guangxi University. His research interests include speech synthesis and natural language processing.

NING WU received the B.E. degree in electrical engineering from Beijing Jiaotong University, in 2003, and the Ph.D. degree in optical engineering from Loughborough University, U.K., in 2007. He is currently a Professor with the College of Electronics and Information Engineering, Beibu Gulf University, China. His research interests include machine learning, image processing, pattern recognition, and holographic microscopy.

CHENGJUN QIU received the B.E. and M.E. degrees from Heilongjiang University, in 1987 and 1993, respectively, and the Ph.D. degree from Harbin Engineering University, in 2005. He is currently a Professor with the School of Electronics and Information Engineering, Beibu Gulf University, China. His research interests include electronic information engineering and machine learning.

YANGYANG LUO received the B.E. degree from Guangxi University, in 2021, where she is currently pursuing the master's degree with the School of Computer and Electronic Information. Her research interests include deep learning and semantic segmentation.

YAN CHEN received the B.E. and M.E. degrees from Guangxi University, China, and the Ph.D. degree from the South China University of Technology, in 2019. She is currently a Professor with the School of Computer and Electronic Information, Guangxi University. Her research interests include machine learning and natural language processing.