applied sciences

IMPACT FACTOR 2.5

CITESCORE 5.5

Article

MixDiff-TTS: Mixture Alignment and Diffusion Model for Text-to-Speech

Yongqiu Long, Kai Yang, Yuan Ma and Ying Yang

MDPI

https://doi.org/10.3390/app15094810

applied sciences

MDPI

Article

MixDiff-TTS: Mixture Alignment and Diffusion Model for Text-to-Speech

Yongqiu Long, Kai Yang *, Yuan Ma and Ying Yang

School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China; 2213393029@st.gxu.edu.cn (Y.L.); li394920l@163.com (Y.M.); yingy2004@126.com (Y.Y.)
* Correspondence: yangkai@gxu.edu.cn

Abstract: In recent years, deep-learning-based speech synthesis has garnered substantial attention, achieving remarkable advancements in generating human-like speech. However, synthesized speech often lacks naturalness, primarily because models excessively depend on fine-grained text–speech alignment. To address this issue, we propose MixDiff-TTS, a novel non-autoregressive model. MixDiff-TTS incorporates a linguistic encoder based on a mixture alignment mechanism, which combines word-level hard alignment with phoneme-level soft alignment. This design reduces reliance on fine-grained alignment, enabling the model to handle ambiguous phonetic boundaries more robustly. Additionally, we introduce a Word-to-Phoneme Attention module with a relative position bias mechanism to improve the model's capacity for processing long text sequences. We evaluate the performance of MixDiff-TTS on the LJSpeech dataset. The experimental results show that MixDiff-TTS scores 0.507 for SSIM (Structural Similarity Index) and 6.652 for MCD (Mel Cepstral Distortion). This suggests that the synthesized speech is closer to real speech in spectral structure and exhibits lower spectral distortion than state-of-the-art baselines (such as FastSpeech2 and DiffSpeech). MixDiff-TTS also achieves a MOS (Mean Opinion Score) of 3.95, which is close to that of real speech. These results indicate that MixDiff-TTS can synthesize speech with high naturalness and quality. Ablation studies demonstrate the effectiveness of our method.

Keywords: deep learning; speech synthesis; non-autoregressive model; mixture alignment

Academic Editor: Douglas O'Shaughnessy

Received: 21 March 2025
Revised: 22 April 2025
Accepted: 23 April 2025
Published: 26 April 2025

Citation: Long, Y.; Yang, K.; Ma, Y.; Yang, Y. MixDiff-TTS: Mixture Alignment and Diffusion Model for Text-to-Speech. *Appl. Sci.* **2025**, *15*, 4810. https://doi.org/10.3390/app15094810

1. Introduction

Speech Synthesis, or Text-to-Speech (TTS), stands as a core foundational technology in artificial intelligence, enabling the seamless transformation of textual content into natural, human-like speech. At its core, it aims to enable machines to communicate seamlessly and naturally with users through synthesized speech [1–3]. This technology serves as a core foundation for human-computer interaction systems, playing a pivotal role in intelligent voice applications. Moreover, it has been deeply integrated into diverse real-world scenarios, including in-vehicle navigation systems, assistive reading devices for the visually impaired, and smart speakers. A standard TTS system architecture is composed of three core modules: a text analysis frontend for linguistic processing, an acoustic model for spectral feature generation, and a vocoder for waveform synthesis [4]. Each module operates within a standardized processing pipeline, utilizing acoustic techniques to extract acoustic features and synthesize speech waveforms. First, the text analysis frontend normalizes raw text through tasks such as word segmentation and polyphonic disambiguation, generating structured linguistic representations including phoneme sequences and prosodic annotations. Next, the acoustic model utilizes deep learning algorithms to map abstract linguistic

features to continuous acoustic representations (e.g., mel-spectrograms), facilitating the transformation from symbolic language to the acoustic domain. Finally, the vocoder employs waveform synthesis techniques to convert acoustic features into high-fidelity speech waveforms, enabling natural-sounding audible output [5,6]. Driven by the rapid advancements in deep learning and the proliferation of human-computer interaction scenarios, neural network-based TTS research has garnered significant attention and achieved remarkable progress [7–10]. During this period, many excellent speech synthesis models have emerged, including WaveNet [11], FastSpeech2 [12], and Glow-TTS [13]. These models have made significant progress in generating natural and fluent human-like speech.

Current TTS models excel in speech synthesis but face limitations in phoneme alignment. For instance, the FastSpeech series models employ explicit phoneme-level hard alignment, which relies on precise boundary annotations [12,14]. However, these annotations are often difficult to obtain. The Tacotron series models adopt an implicit alignment method based on attention mechanisms [15,16]. This type of alignment can complete phoneme alignment without external annotations, but the model is prone to alignment instability issues during training, such as skips and repetitions. The instability of these alignment methods weakens the expressiveness of synthetic speech. Additionally, the recently proposed DiffSpeech [17] model enables efficient training by optimizing the variational lower bound, generating highly realistic mel-spectrograms that closely match real data distributions. It also enables the synthesis of speech featuring sophisticated variations in expressiveness. However, DiffSpeech uses phoneme-level hard alignment during phoneme alignment, which sometimes results in unnatural-sounding speech. This is mainly because phoneme boundaries are inherently uncertain. For models relying on phoneme-level hard alignment, the phoneme boundaries obtained through this process often lack sufficient precision to meet alignment requirements. Consequently, ambiguous phoneme boundaries during alignment inevitably introduce errors, which degrade the prosodic naturalness of speech synthesis.

To address the problem of phoneme sequence alignment and improve the quality of synthesized speech, we propose a non-autoregressive model called MixDiff-TTS. Inspired by advanced TTS models [18], we introduce a linguistic encoder in MixDiff-TTS. To achieve fine-grained phoneme alignment, this encoder innovatively incorporates a mixture alignment mechanism. This mechanism enables the model to preserve semantic boundary information while more effectively modeling ambiguous boundary regions. The structure of this linguistic encoder effectively solves the issue of phoneme-level hard alignment. To improve the model's ability to handle long text sequences, we introduce a Word-to-Phoneme Attention module with a relative position bias mechanism. We incorporate a pre-net structure to enhance the model's learning capability for speech synthesis tasks. Additionally, to further enhance mel-spectrogram reconstruction quality, we incorporate a post-net. Finally, the mel-spectrogram processed by the post-net is converted into a speech waveform using the HiFi-GAN [19] vocoder. We conduct comprehensive evaluations of MixDiff-TTS on the LJSpeech dataset. The experimental results show that MixDiff-TTS outperforms other TTS models in evaluations of SSIM, MCD, and MOS, demonstrating its superiority in speech quality, naturalness, and intelligibility.

The main contributions of this paper are as follows:

1. We propose MixDiff-TTS, a non-autoregressive TTS model that integrates a mixture alignment mechanism with a diffusion model.
2. We introduce a Word-to-Phoneme Attention module with relative position bias to improve the model's ability to handle long text sequences.
3. We incorporate a pre-net structure to enhance the model's learning capability for speech synthesis tasks.

4. We incorporate a post-net structure to optimize the reconstruction quality of mel-spectrograms.
5. Our objective and subjective evaluations show that MixDiff-TTS outperforms baselines in multiple metrics, validating its effectiveness.

2. Materials and Methods

The following section is divided into two subsections. The first subsection introduces the background material required for this paper, and the second subsection describes our model.

2.1. Background

2.1.1. Diffusion Models

Owing to their exceptional capability for modeling data, diffusion models are widely employed to characterize complex data distributions. As a subclass of probabilistic generative models, they have achieved cutting-edge performance in diverse domains (e.g., image and audio synthesis) in recent years, thereby demonstrating their remarkable capacity to produce high-fidelity samples [20–25]. Inspired by non-equilibrium thermodynamics in physics, diffusion models employ a bidirectional stochastic process framework grounded in Markov chains. During training, the diffusion process gradually injects Gaussian noise into data samples, thereby progressively degrading their structural coherence until they evolve into pure Gaussian noise. Subsequently, the models achieve sample reconstruction by learning a reverse process, thereby recovering high-fidelity original samples from this noise-perturbed state. As depicted in Figure 1, the bidirectional stochastic process of diffusion models is illustrated.

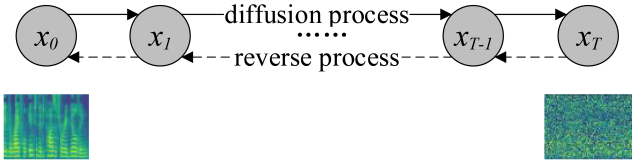


Figure 1. Schematic diagram of the diffusion process and reverse process.

Given the predefined noise schedule β and diffusion steps t, the constants corresponding to the diffusion process and the reverse process are calculated, as shown in the following equation:

$$\alpha_t = \prod_{i=1}^{t} \sqrt{1-\beta_i} \tag{1}$$

$$\sigma_t = \sqrt{1-\alpha_t^2} \tag{2}$$

where β represents the noise schedule that governs the magnitude of noise injected during the diffusion process. α_t denotes the cumulative retention rate at each diffusion step t, measuring the signal attenuation from the original data to the current diffusion step. σ_t is the total noise standard deviation at diffusion step t, indicating the proportion of noise in the current state.

The diffusion process is typically defined mathematically by a memoryless Markov chain [26]. This Markov chain is characterized by two key properties: its parameters remain constant across all diffusion steps, and the current state depends solely on the previous state. The properties of the Markov chain significantly reduce the computational complexity

of the diffusion process. By gradually adding Gaussian noise to the initial data x_0, the diffusion process can transform x_0 into a latent variable x_T over T steps, as shown below:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3}$$

where $q(x_{1:T}|x_0)$ represents the joint probability distribution of all intermediate variables x_1 to x_T in the diffusion process, given that x_0. $q(x_t|x_{t-1})$ denotes the conditional probability of a single diffusion step, describing how x_t is generated from x_{t-1} in the diffusion process.

According to the predefined noise schedule β, the diffusion process can obtain the latent variable x_T by adding small Gaussian noise to x_{t-1}. Therefore, $q(x_t|x_{t-1})$ is given as follows:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{4}$$

where I denotes the identity matrix, ensuring that the noise is isotropic (i.e., independent across dimensions). This equation indicates that each step x_t is sampled from a Gaussian distribution with a mean of $\sqrt{1-\beta_t}x_{t-1}$ and a variance of $\beta_t I$.

The reverse process is also modeled by a Markov chain. Unlike the diffusion process with fixed parameters, the parameters θ of this Markov chain are learned through training. By removing the added noise, data x_0 are gradually generated from the latent variable x_T, as shown in the following equation:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{5}$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{6}$$

where $p_\theta(x_{t-1}|x_t)$ represents the single-step transition probability in the reverse process, describing how x_t generates the previous diffusion step x_{t-1}. $\mu_\theta(x_t, t)$ and $\sigma_t^2 I$ respectively denote the conditional mean and variance of the process. The joint probability distribution of the entire reverse process is given by $p_\theta(x_{0:T})$. This indicates that starting from the final pure Gaussian noise x_T, we can progressively denoise through a Markov chain in the reverse direction to generate a clean data x_0. The data distribution $p(x_T)$ at the final diffusion step is typically a standard Gaussian distribution $\mathcal{N}(0, I)$.

The training objective of diffusion models is to learn parameters θ to generate samples that match the real data distribution from noise. Therefore, we train the model by minimizing the variational lower bound of the negative log-likelihood and optimize it using stochastic gradient descent [26]. The computation is detailed as follows:

$$E_{q(x_0)}[-\log p_\theta(x_0)] \geq E_{q(x_0, x_1, \ldots, x_T)}[\log q(x_{1:T}|x_0) - \log p_\theta(x_{0:T})] = L \tag{7}$$

$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \tag{8}$$

where D_{KL} denotes the Kullback–Leibler (KL) divergence, which is a metric used to quantify the difference between two probability distributions.

2.1.2. HiFi-GAN Vocoder

A vocoder is a system used for analyzing and synthesizing speech signals. As a key component of speech synthesis systems, its performance directly determines the naturalness and clarity of synthesized speech. Early neural vocoders, such as WaveNet [11] and WaveGlow [27], can convert the acoustic features predicted by acoustic models into audio waveforms. However, these vocoders have several issues [28,29]. First, WaveNet generates audio in an autoregressive manner (i.e., sequentially by time steps). Its sequential

generation nature leads to slow inference speed and low real-time performance. Second, the core of WaveGlow is invertible neural networks, which directly model data distributions through latent variable transformations. However, its complex structural design results in high training costs.

To address these issues, Jungil Kong et al. proposed the HiFi-GAN vocoder based on the Generative Adversarial Network (GAN). Its architecture consists of two core components: a generator and a discriminator. The generator uses transposed convolutions to perform multi-scale upsampling on the mel-spectrogram and innovatively introduces the Multi-Receptive Field Fusion (MRF) module. The MRF module captures multi-time-scale features of the speech signal by stacking residual blocks with different convolutional kernel sizes in parallel. The discriminator is composed of a Multi-Scale Discriminator (MSD) and a Multi-Period Discriminator (MPD). The MSD extracts features at different scales through multiple average pooling operations. The MPD uses two-dimensional convolutions to capture the periodic features of the audio. Leveraging the unique structures of the generator and the discriminator, HiFi-GAN can generate high-fidelity speech in real time. Compared with early neural vocoders, HiFi-GAN has three significant advantages [19]. First, it achieves efficient inference through a non-autoregressive architecture. In terms of inference speed, HiFi-GAN is approximately seven times faster than WaveGlow and significantly outperforms WaveNet. Second, it adopts a lightweight network structure. The number of parameters in HiFi-GAN is only half that of WaveNet and one-sixth that of WaveGlow, reducing the demand for computational resources. Third, through the MRF module and dual discriminators, HiFi-GAN can generate high-fidelity speech. In the MOS evaluation, HiFi-GAN improves by 0.34 points compared to WaveNet and 0.55 points compared to WaveGlow.

Additionally, as research on diffusion models has deepened, a diffusion-model-based vocoder has emerged (e.g., DiffWave [30]). This type of vocoder gradually converts random noise into high-quality audio by introducing diffusion models. However, since generating audio with diffusion models requires dozens to hundreds of iterations, their inference speed is slow. In contrast, HiFi-GAN can map all input features into an audio waveform at once, so it has high real-time performance [31]. In real-time speech synthesis scenarios, HiFi-GAN leverages the efficient architecture of the MRF module to generate high-fidelity speech comparable to diffusion-based vocoders, without sacrificing inference speed. Therefore, we select HiFi-GAN as the vocoder for MixDiff-TTS.

2.2. MixDiff-TTS

2.2.1. Motivation

Although previous models (e.g., DiffSpeech [17] and FastSpeech2 [12]) have demonstrated great potential in synthesizing high-quality speech samples, there are still limitations. In these models, phoneme durations are typically predicted by a duration predictor, while the real phoneme durations are derived from the Montreal Forced Aligner (MFA). However, using MFA to obtain phoneme durations leads to the problem of phoneme ambiguity during alignment. This issue primarily stems from the ill-defined boundaries between phonemes. Therefore, to address alignment issues arising from ambiguous phoneme boundaries in phoneme-level hard alignment, MixDiff-TTS introduces a linguistic encoder that incorporates a mixture alignment mechanism. The model utilizes this mechanism to achieve soft alignment at the phoneme level while retaining hard alignment at the word level. Specifically, the mixture alignment mechanism stabilizes the training process by introducing structural priors (such as hard alignment) while retaining the fine-grained expressive power of the soft alignment mechanism, thus achieving a good balance between naturalness and stability. Building on this, to enhance the model's ability to handle long

text sequences, MixDiff-TTS introduces a Word-to-Phoneme Attention module with relative position bias. Additionally, MixDiff-TTS incorporates a pre-net structure to improve the model's learning ability for speech synthesis tasks.

In speech synthesis tasks, the mel-spectrograms can help the model capture and generate the key characteristics of speech. Moreover, the quality of synthesized speech is substantially influenced by the representational capacity of these spectrograms. For speech synthesis, optimizing the reconstruction ability of the mel-spectrograms is a crucial step. Therefore, we enhance their feature representation ability by optimizing the reconstruction capability of the mel-spectrograms. Specifically, we enhance the mel-spectrograms' capacity for modeling fine-grained acoustic details by introducing a post-net into MixDiff-TTS, thereby improving their reconstruction ability. Finally, ablation studies demonstrate that each newly introduced component effectively improves the speech synthesis performance of MixDiff-TTS. Figure 2a illustrates the detailed architecture of MixDiff-TTS.

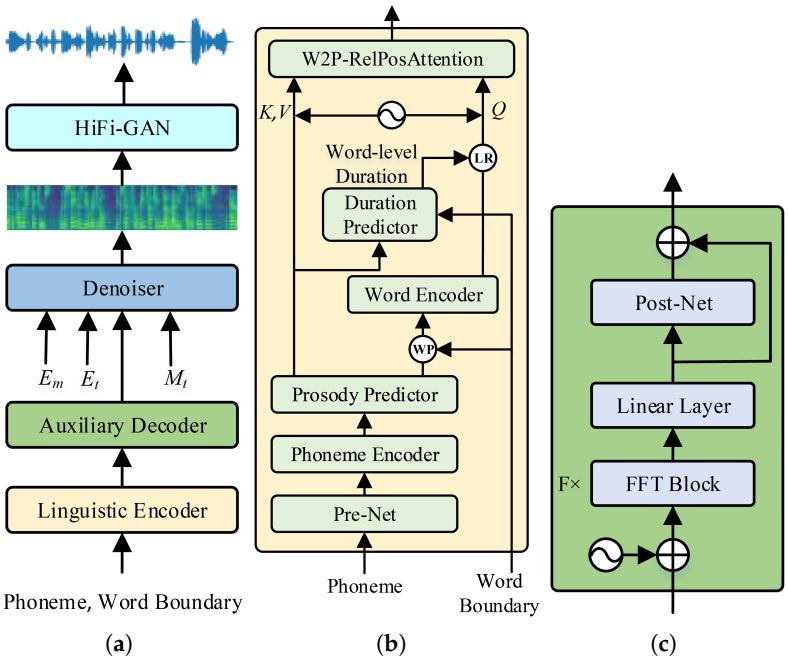


Figure 2. Architecture diagram of MixDiff-TTS and its modules. (**a**) Overall architecture of MixDiff-TTS. M_t is the mel-spectrogram at the t-th step in the diffusion process, E_t is the step embedding, and E_m is the sequence output by the linguistic encoder. (**b**) Detailed structure of the Linguistic Encoder. (**c**) Detailed structure of the Auxiliary Decoder.

2.2.2. Linguistic Encoder

Figure 2b depicts the detailed architecture of the linguistic encoder, where "LR" refers to the length regulator for adjusting input sequence length, and "WP" signifies the transformation of input phoneme representations into word representations through word-level pooling. The pre-net structure performs multi-layer nonlinear transformations on the input phoneme sequence. By using activation functions and a dropout mechanism, the model's capacity for feature representation is significantly enhanced. Specifically, its architecture includes multiple fully connected layers and ReLU activations. This structure captures the potential nonlinear relationships between phonemes by performing nonlinear mapping on the original phoneme embeddings. With this design, the model becomes more effective in modeling the phoneme-to-acoustic feature mapping, thereby further enhancing its learning capacity. The phoneme encoder and word encoder share identical components, as depicted in Figure 3a. Specifically, both are constructed by stacking multiple Feed-Forward Transformer (FFT) layers. By stacking FFT layers, they effectively capture sequence features

from both local and global contexts, and separately model feature representations at the phoneme and word levels. As shown in Figure 3b, the prosody predictor consists of a pitch predictor and an energy predictor. We employ architectures for both that are identical to the corresponding modules in FastSpeech2 [12], as depicted in Figure 3c. These two predictors, respectively, predict the pitch and energy features in speech. Additionally, prosody prediction is also influenced by duration as pitch and duration are closely related. During prosody prediction, stressed segments tend to have longer durations and may also exhibit higher pitch. The phoneme sequence and word-level hidden states output by the linguistic encoder serve as the common foundation for both pitch and duration prediction.

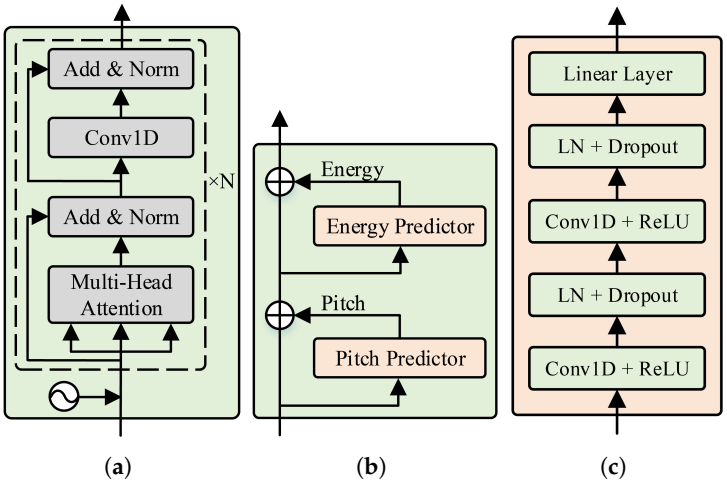


Figure 3. (**a**) Detailed structure of the Phoneme/Word Encoder. (**b**) Detailed structure of the Prosody Predictor. (**c**) Detailed structure of the Pitch/Energy Predictor.

The W2P-RelPosAttention module is a Word-to-Phoneme Attention module with a relative position bias mechanism. In this module, to enable the query Q to find the phonemes associated with the corresponding word, an additional mapping mask is introduced into the attention weights. Building on this, we introduce a relative position bias mechanism to enhance the model's capability for modeling long text sequences, with its computation defined as follows:

$$output = Softmax(\frac{QK^T}{\sqrt{d_k}} + P(i,j))V + Q \tag{9}$$

where i and j are the indices of elements in the query Q and the key K, respectively. d_k denotes the dimension of K. $P(i,j)$ is the relative position bias, representing the relative position information between Q and K. $\frac{QK^T}{\sqrt{d_k}}$ is used to calculate the similarity between Q and K. The purpose of the expression $\frac{QK^T}{\sqrt{d_k}} + P(i,j)$ is to enable the model to capture the relative positional relationship between Q and K, rather than merely focusing on their content-based matching similarity. Through encoding the relative positional relationship between Q and K, this mechanism adjusts the attention scores to incorporate positional dependencies. As a result, the model can make adaptive adjustments across different parts of the sequence. When dealing with long text sequences, the model can understand the positional relationships based on the specific context instead of simply relying on the fixed sequence positions. This enables the model to more effectively capture dependencies in long text sequences.

2.2.3. Auxiliary Decoder

We introduce a mel-spectrogram decoder named the auxiliary decoder, whose architecture is shown in Figure 2c. This decoder primarily consists of stacked FFT blocks. In

each FFT block, the number of layers F is 4, the self-attention hidden size is 256, the number of attention heads is 2, and the kernel sizes of the two 1D convolutional layers are 9 and 1, respectively. Within the FFT blocks, the sequence's contextual information is captured using the attention mechanism. After passing through multiple FFT layers, the output is a sequence enriched with contextual features. Subsequently, the linear layer processes the output from the FFT module to generate prediction results. However, this output is a coarse mel-spectrogram with limited feature representation capability. Therefore, to address the limitations in mel-spectrogram reconstruction, we incorporate a post-net into the auxiliary decoder. Figure 4 illustrates the detailed architecture of the post-net. It is a five-layer 1D convolutional network. Each layer includes 512 5 × 1 convolutional kernels, batch normalization, and dropout. It enhances the ability to restore details in the prediction results by capturing the residual features of the mel-spectrograms. Compared with 2D convolutions or recurrent network structures, 1D convolutions are more suitable for processing the temporal dimension features of spectrograms. They can extract local contextual information while maintaining computational efficiency. Multi-layer stacked 1D convolutional networks can gradually expand the receptive field to capture structural information over longer time spans, thereby effectively refining the prediction results. After processing by the post-net, the optimized mel-spectrograms are obtained.

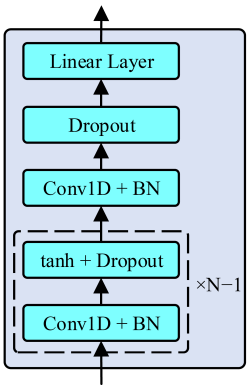


Figure 4. Detailed structure of the post-net.

2.2.4. Denoiser

Similar to DiffSpeech [17], we adopt the non-causal WaveNet architecture as the denoiser. The non-causal WaveNet architecture leverages contextual information and parallel prediction to directly learn multi-scale features from raw audio. Moreover, this architecture improves the continuity and naturalness of denoised signals by minimizing regression losses (e.g., L1 loss) [32]. Due to its efficient parallel computing and mature audio generation capabilities, it has become an ideal choice for denoisers in diffusion-based speech synthesis [9,17].

The denoiser takes M_t as input and is conditioned on E_t and E_m to predict the noise added during the diffusion process. It comprises a 1 × 1 convolutional layer that projects M_t into a hidden sequence and N convolutional blocks. Each convolutional block consists of five components:

1. An element-wise addition operation to add E_t to the hidden sequence.
2. A non-causal convolutional layer that transforms the hidden sequence from C to $2C$ channels (C typically set to 256).
3. A 1 × 1 convolutional layer mapping E_m to $2C$ channels.
4. A gating unit that fuses the input information and conditional information.

5. A residual block that splits the fused hidden states into two branches, each with C channels. This structure enables the denoiser to fuse features across hierarchical levels, thereby generating the final prediction.

2.2.5. Training Loss

The training of MixDiff-TTS is divided into two stages. First, an auxiliary decoder is trained, and then its outputs serve as conditions for training the diffusion model. Therefore, the total training loss comprises the auxiliary decoder's training loss L_{aux} and the diffusion model's training loss L_{diff}:

$$L_{aux} = L_{mel} + L_{duration} + L_{pitch} + L_{energy} + L_{helper} \tag{10}$$

$$L_{diff} = L_{noise} + L_{duration} + L_{pitch} + L_{energy} \tag{11}$$

where L_{mel} employs Mean Absolute Error (MAE) to compare predicted and real mel-spectrograms. $L_{duration}$, L_{pitch}, and L_{energy} use Mean Squared Error (MSE) to calculate differences between predicted and real values. L_{helper} uses the guided attention loss [33], and L_{noise} uses MAE to compute the error between the predicted noise and the denoised output.

3. Results

To evaluate the speech synthesized by MixDiff-TTS, we conduct comparative experiments with FastSpeech2 [12], PortaSpeech [18], and DiffSpeech [17] as baselines. This section is structured into three parts. The first subsection describes the experimental datasets. The second subsection details the configuration of MixDiff-TTS. The third subsection describes the evaluation experiments.

3.1. Dataset

3.1.1. Dataset Selection

We use the LJSpeech dataset for our evaluation experiments. This dataset is a public-domain single-speaker speech corpus that contains 13,100 audio clips. Each audio clip is provided with corresponding transcribed text. The duration of the clips ranges from 1 to 10 s. These clips total approximately 24 h of audio.

3.1.2. Dataset Processing

We split the LJSpeech dataset into training, validation, and test sets, containing 12,076, 512, and 512 samples, respectively. For the evaluation, we randomly select 50 samples from the test set for subjective evaluation, while all test samples are used for objective evaluation. All audio clips are sampled at 22,050 Hz. Additionally, we set the frame size to 1024 and the hop length to 256.

3.2. Model Configuration

MixDiff-TTS is trained on an NVIDIA 4090 GPU. We utilize the Adam [34] optimizer for model optimization, with hyperparameters configured as $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The model adopts the learning rate scheduling and diffusion step configuration described in [17]. MixDiff-TTS achieves high-quality synthesized speech through 100 diffusion steps. In the first training stage, the loss function converges after 300 k steps, whereas the second stage involves 600 k steps. The experimental software environment includes CUDA 12.6, Python 3.8, g2p-en 2.1.0, and PyTorch 2.0.0+cu118. All baseline models are tested using their publicly available GitHub (PortaSpeech v0.2.0) implementations and employ the HiFi-GAN vocoder with the same configuration to synthesize audio waveforms

as the final step. During evaluation, text content is selected following the principle of consistency to control confounding variables.

3.3. Evaluation

3.3.1. Evaluation Metrics

For evaluation, we adopt four objective metrics: SSIM [35], MCD [36], fundamental frequency (F_0) root mean squared error (RMSE), and Fréchet Audio Distance (FAD) [37]. In speech synthesis, SSIM is commonly used to measure the structural similarity between real and synthesized mel-spectrograms. The mathematical formulation of SSIM is as follows:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{12}$$

where x and y are the elements of the real and synthesized spectrograms, respectively. μ_x and μ_y denote their respective means, σ_x^2 and σ_y^2 represent their respective variances, σ_{xy} is their covariance, and C_1 and C_2 are constants that prevent the denominator from becoming zero. An SSIM value closer to 1 signifies greater structural similarity between synthesized and real mel-spectrograms.

MCD is widely adopted to quantify spectral shape distortion between real and synthesized speech. The formula of MCD is shown below:

$$\text{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}(c_m - \hat{c}_m)^2} \tag{13}$$

where M represents the dimensionality of Mel-Cepstral Coefficients (MCCs). Specifically, c_m denotes the m-th dimensional MCC of natural speech, while $\hat{c}_m$ corresponds to the m-th dimensional MCC of synthesized speech. As a key spectral feature in speech processing, MCCs are widely used in academic research to characterize the overall spectral contour of speech. A lower MCD value indicates smaller cepstral distortion, implying that the spectral characteristics of synthesized speech more closely match those of natural speech.

F_0 RMSE is a commonly used objective metric for assessing the accuracy of F_0 predictions. Specifically, its role is to quantify the discrepancy between predicted and true F_0 values. The formula for F_0 RMSE is as follows:

$$F_0\,\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - \hat{F}_i)^2} \tag{14}$$

where F_i denotes the real fundamental frequency of the i-th speech frame, while $\hat{F}_i$ denotes its corresponding predicted fundamental frequency. A smaller F_0 RMSE signifies that the fundamental frequency contours of synthesized and real speech are more closely matched, thereby reflecting more consistent time-domain dynamic characteristics of fundamental frequency.

FAD, which is designed as a metric to quantify the quality of generated audio, measures the similarity of real to generated audio by calculating the distributional distance in their feature spaces. It relies on the VGGish audio feature extractor to map audio into a perceptual feature space and then computes the Fréchet distance between Gaussian distributions in this space. Its calculation formula is as follows:

$$\text{FAD} = \|\mu_a - \mu_b\|^2 + tr(\sum_a + \sum_b - 2\sqrt{\sum_a \sum_b}) \tag{15}$$

where a and b denote the real and synthesized speech, respectively. μ represents the mean vector of speech features, $\sum$ denotes the covariance matrix, and tr indicates the matrix trace. A smaller FAD value implies a closer match of the synthesized speech distribution with that of real speech, thereby indicating a higher degree of similarity between the two.

We adopt two subjective evaluation metrics: MOS [38] and Comparative Mean Opinion Score (CMOS) [39], both of which are based on human ratings. Additionally, the Real-Time Factor (RTF) is employed to measure the model's audio generation speed. An RTF value below 1.0 indicates that the model generates audio quickly enough for real-time applications, and lower RTF values indicate better real-time performance.

3.3.2. Experimental Results

In the evaluation experiments, we use a logarithmic method to calculate the F_0 RMSE. To maintain fairness in assessing the MCD and F_0 RMSE, we employ Dynamic Time Warping (DTW) [40] to establish alignment between synthesized speech and real speech samples. All evaluation results are summarized in Table 1. Compared to baseline models, MixDiff-TTS exhibits leading performance in SSIM and MCD assessments. However, its F_0 RMSE is slightly higher than some baselines, indicating room for improvement in fundamental frequency prediction. These findings indicate that MixDiff-TTS demonstrates satisfactory performance in phoneme alignment and can synthesize high-fidelity audio. In the FAD evaluation, MixDiff-TTS outperforms other models. This demonstrates that its synthesized audio is highly similar to real audio, with their feature distributions closely aligned. "Params" refers to the total number of model parameters. We observe that MixDiff-TTS still has room for improvement in optimizing the parameter count. Additionally, to investigate the real-time synthesis performance of MixDiff-TTS, we selected 10 synthesized audio samples for RTF evaluation. These samples contain between 10 and 20 words. An RTF value less than 1 indicates that MixDiff-TTS has a fast inference speed and high real-time performance.

We set the confidence interval for the MOS evaluation at 95%. To ensure the statistical reliability of subjective evaluation results, each sample is rated by at least 10 participants. All participants wear headphones and complete all evaluations in a quiet environment. Prior to the test, we shuffle the order of all samples to ensure that participants are unaware of which model generated each sample. We instruct participants to rate the audio based on pronunciation accuracy and naturalness using a 5-point scale ranging from 1 to 5, with 0.5-point increments between scale points. As shown in Table 1, MixDiff-TTS achieves a MOS score of 3.95, outperforming FastSpeech2 and DiffSpeech while slightly trailing PortaSpeech. This result demonstrates that MixDiff-TTS effectively models phoneme alignment and can synthesize audio with high pronunciation accuracy and naturalness. Furthermore, we use the CMOS metric to compare the performance among the models involved in the test. With FastSpeech2 as the baseline, participants evaluate speech samples generated by comparative models. As can be observed, MixDiff-TTS performs comparably to PortaSpeech and outperforms DiffSpeech, indicating that the model can generate higher-quality synthesized speech.

Table 1. Results of objective and subjective evaluations and model efficiency tests.

Model	SSIM	MCD	F_0 RMSE	FAD	Params	RTF	MOS	CMOS
Ground Truth							4.37 ± 0.08	
FastSpeech2	0.496	6.751	0.329	2.1460	35.16M	0.1486	3.82 ± 0.09	0.000
PortaSpeech	0.505	6.683	0.323	1.8593	23.97M	0.1364	3.98 ± 0.07	0.189
DiffSpeech	0.501	6.735	0.335	1.5165	45.11M	0.1456	3.92 ± 0.08	0.179
MixDiff-TTS	0.507	6.652	0.337	1.5013	46.31M	0.1391	3.95 ± 0.08	0.185

3.4. Feature Visualization

We explore the prediction of mel-spectrograms and the visualization of attention alignment. In speech synthesis, mel-spectrograms featuring clear harmonic structures and broad frequency coverage directly signify superior phonetic clarity and prosodic naturalness in synthesized speech. Thus, we perform a comparative analysis of mel-spectrograms generated by all tested models across three key dimensions: adjacent harmonic structures, unvoiced frame distributions, and low-frequency region characteristics. In mel-spectrograms, adjacent harmonics typically manifest as parallel bright bands, which correspond to the frequency distribution of the fundamental frequency and its harmonic components. A well-defined harmonic structure generally indicates pitch stability. Unvoiced frames usually appear as dark regions in the mel-spectrogram, where no prominent harmonics are present. The low-frequency region is typically located in the lower part of the mel-spectrogram. Figure 5 shows the mel-spectrograms generated by all tested models. The horizontal axis denotes the frame number, with each frame typically corresponding to an audio segment of approximately 10 ms. The vertical axis denotes the frequency in the mel-spectrogram, with values ranging from 0 to 80 indicating 80 mel-frequency bins. The spectrograms within the black boxes show that the mel-spectrograms synthesized by MixDiff-TTS exhibit a clearer harmonic structure, more accurate unvoiced frame prediction, and a more stable low-frequency region. Experimental results demonstrate that the MixDiff-TTS model has significant advantages in generating detailed mel-spectrogram characteristics.

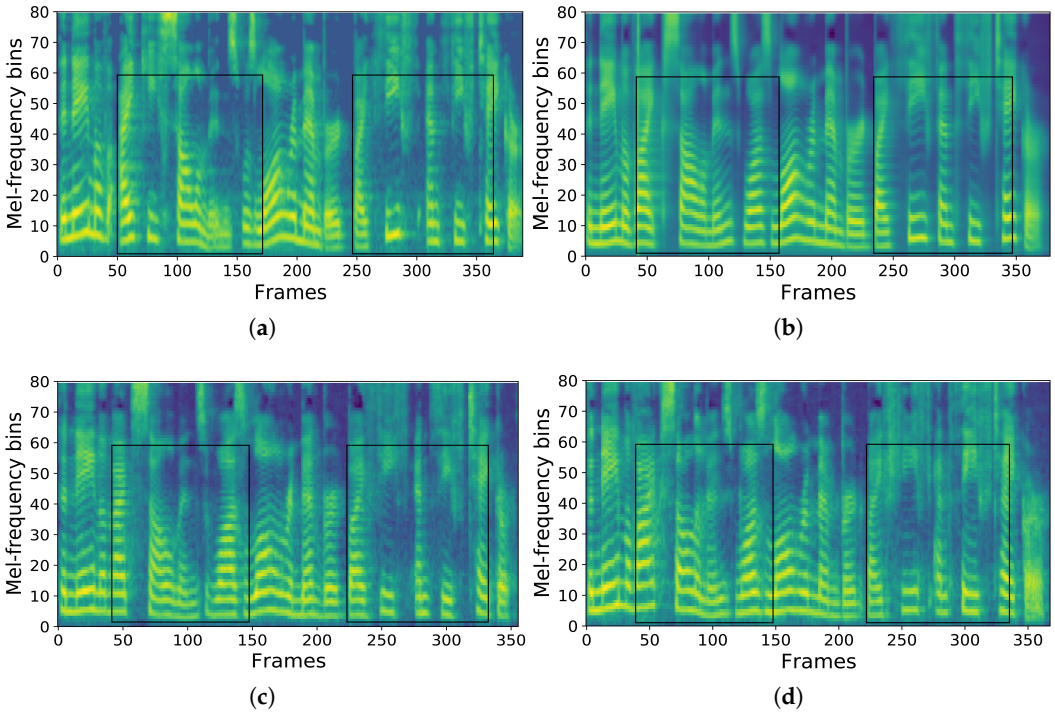
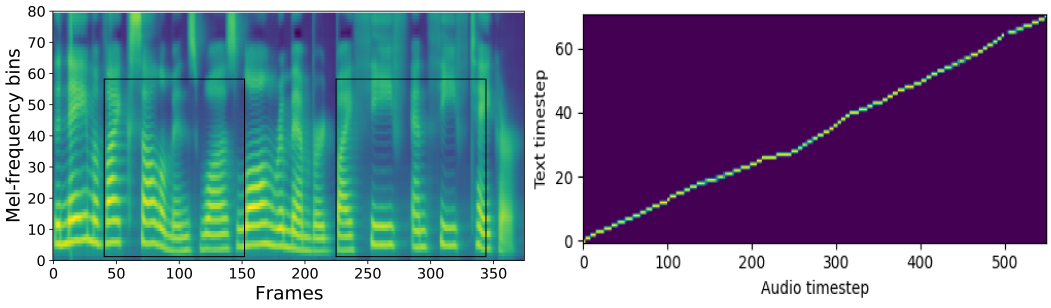


Figure 5. *Cont.*



(**e**) (**f**)

Figure 5. Comparison of mel-spectrograms from different TTS models and attention alignment of MixDiff-TTS. (**a**) Ground-truth mel-spectrogram. (**b**) Mel-spectrogram predicted by FastSpeech2. (**c**) Mel-spectrogram predicted by PortaSpeech. (**d**) Mel-spectrogram predicted by DiffSpeech. (**e**) Mel-spectrogram predicted by MixDiff-TTS. (**f**) Attention alignment of MixDiff-TTS.

Figure 5f illustrates the attention alignment visualization of MixDiff-TTS. The horizontal axis denotes audio time steps, with each step corresponding to an audio segment of approximately 10 ms. The vertical axis denotes text time steps, with each unit corresponding to a token (e.g., a phoneme) in the sequence. In the attention alignment map, dots denote attention weights, with brighter points indicating higher values. The dot-formed diagonal represents the alignment of phoneme and mel-spectrogram sequences, where a clearer diagonal signifies higher alignment accuracy. In Figure 5f, the dots exhibit distinct brightness, and the diagonal line is clear and smooth, demonstrating the significant capability of MixDiff-TTS in aligning phoneme sequences with mel-spectrograms.

3.5. Ablation Studies

To investigate the effectiveness of the mixture alignment scheme, we conduct comparative experiments between it and phoneme-level hard alignment methods. "MixDiff-TTS—phoneme-level hard alignment" indicates that MixDiff-TTS uses phoneme-level hard alignment. In Table 2, MixDiff-TTS with the mixture alignment mechanism demonstrates superior performance over its phoneme-level hard alignment counterpart in both MOS and CMOS scores, thereby indicating that this mechanism significantly enhances the prosodic naturalness of synthesized speech.

Table 2. The ablation study on mixture alignment.

Setting	MOS	CMOS
MixDiff-TTS	3.95 ± 0.08	0.000
MixDiff-TTS—phoneme-level hard alignment	3.83 ± 0.07	-0.247

To evaluate the impact of the MixDiff-TTS architecture on speech generation, we conduct ablation studies on the W2P-RelPosAttention module, pre-net, post-net, and residual network. Each ablation study is evaluated based on SSIM, MCD, F_0 RMSE, MOS, and CMOS. "MixDiff-TTS—W2P-RelPosAttention" refers to the MixDiff-TTS model after removing the W2P-RelPosAttention module. In Table 3, MixDiff-TTS with the W2P-RelPosAttention module achieves the highest SSIM and lowest MCD scores, whereas its F_0 RMSE performance remains suboptimal. The MOS and CMOS results indicate that MixDiff-TTS outperforms its variant without the W2P-RelPosAttention module. The experimental results suggest that MixDiff-TTS becomes more competitive in handling long-sequence text after incorporating the W2P-RelPosAttention module.

Table 3. Ablation study of the W2P-RelPosAttention module.

Setting	SSIM	MCD	F_0 RMSE	MOS	CMOS
MixDiff-TTS	0.507	6.652	0.337	3.95 ± 0.08	0.000
MixDiff-TTS—W2P-RelPosAttention	0.503	6.679	0.329	3.91 ± 0.08	−0.135

"MixDiff-TTS—pre-net" refers to the MixDiff-TTS model after removing the pre-net structure. As shown in Table 4, removing the pre-net results in degraded performance in SSIM, MCD, F_0 RMSE, MOS, and CMOS for MixDiff-TTS. Experimental results demonstrate that the pre-net significantly enhances the model's capacity for learning the speech synthesis task. In Table 5, "MixDiff-TTS—residual network" denotes the MixDiff-TTS variant that directly outputs post-net processed results after removing the residual network. "MixDiff-TTS—post-net" denotes the MixDiff-TTS model after removing the post-net structure. The findings indicate that MixDiff-TTS with the post-net and residual network achieves the best performance in SSIM and MCD, although it does not show an improvement in F_0 RMSE. In subjective evaluation, MixDiff-TTS without the post-net and residual network exhibits significantly poorer performance than the baseline model retaining these components. These experimental findings demonstrate that the inclusion of the post-net significantly boosts MixDiff-TTS's capacity for mel-spectrogram reconstruction.

Table 4. The ablation study on the pre-net.

Setting	SSIM	MCD	F_0 RMSE	MOS	CMOS
MixDiff-TTS	0.507	6.652	0.337	3.95 ± 0.08	0.000
MixDiff-TTS—pre-net	0.501	6.661	0.372	3.93 ± 0.07	−0.129

Table 5. Ablation study of the post-net and residual network.

Setting	SSIM	MCD	F_0 RMSE	MOS	CMOS
MixDiff-TTS	0.507	6.652	0.337	3.95 ± 0.08	0.000
MixDiff-TTS—residual network	0.501	6.725	0.345	3.92 ± 0.07	−0.131
MixDiff-TTS—post-net	0.505	6.793	0.331	3.86 ± 0.08	−0.218

4. Discussion

In the evaluation experiments, we observed that MixDiff-TTS has higher F_0 RMSE values compared with other TTS models. We believe that this may be attributed to the inherent randomness within the generation process of the diffusion model. Although this randomness helps enhance the diversity and naturalness of generated speech, it may also cause slight deviations in the F_0 trajectory of synthesized speech compared to real speech. This slight deviation may lead to a higher F_0 RMSE. Therefore, it is necessary to further optimize the prosody modeling mechanism in future work to enhance the model's synthesis accuracy in the F_0 dimension.

Additionally, while MixDiff-TTS can synthesize high-quality speech, it has design limitations:

1. In speech synthesis tasks, multi-speaker speech synthesis models have become a research hotspot to meet the needs of different scenarios. Such models can generate speech with different genders and timbres without retraining for each specific requirement, making them suitable for various applications. However, MixDiff-TTS is only applicable to single-speaker speech synthesis and lacks support for multi-speaker tasks, limiting its applicability in broader scenarios;

2. In speech synthesis, fully end-to-end TTS models have emerged as a pivotal research focus, gaining substantial attention in recent years. A fully end-to-end model can directly generate speech waveforms from raw text using a unified architecture, without relying on intermediate feature representations (such as mel-spectrograms). In contrast, MixDiff-TTS generates intermediate feature representations during synthesis and then uses a vocoder to convert them into final speech waveforms. This results in the need for separate training of the acoustic model and vocoder, incurring additional training costs.

Therefore, based on MixDiff-TTS, developing a fully end-to-end multi-speaker speech synthesis model is necessary.

5. Conclusions

In this paper, we propose MixDiff-TTS, a non-autoregressive speech synthesis model. MixDiff-TTS introduces a linguistic encoder with a mixture alignment mechanism to address the phoneme-level hard alignment issue. Building on this, MixDiff-TTS incorporates a Word-to-Phoneme Attention module with relative position bias to enhance the model's capacity for processing long text sequences. To enhance the model's learning capability for the speech synthesis task, a pre-net structure is introduced. Additionally, a post-net structure is incorporated to enhance the reconstruction quality of mel-spectrograms.

We perform both objective and subjective evaluations of MixDiff-TTS on the LJSpeech dataset. The experimental results demonstrate that MixDiff-TTS exhibits strong competitiveness in synthesizing high-quality speech compared to other baseline models. Furthermore, to validate the necessity of each component, we conduct ablation studies on the components integrated into MixDiff-TTS. In future work, we aim to build on our current research to develop a fully end-to-end multi-speaker TTS model.

Author Contributions: Conceptualization, K.Y.; methodology, Y.L.; software, Y.L.; validation, Y.L.; formal analysis, K.Y.; investigation, Y.M.; writing—original draft preparation, Y.L.; writing—review and editing, K.Y. and Y.Y.; supervision, K.Y.; project administration, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The LJSpeech dataset is available at https://keithito.com/LJ-Speech-Dataset (accessed on 21 March 2025).

Acknowledgments: I sincerely appreciate each reviewer for their insightful and professional feedback on my manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, J.; Xie, Z.; Zhang, C.; Shi, G. A novel method for Mandarin speech synthesis by inserting prosodic structure prediction into Tacotron2. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 2809–2823. [CrossRef]
2. de Barcelos Silva, A.; Gomes, M.M.; Da Costa, C.A.; da Rosa Righi, R.; Barbosa, J.L.V.; Pessin, G.; De Doncker, G.; Federizzi, G. Intelligent personal assistants: A systematic literature review. *Expert Syst. Appl.* **2020**, *147*, 113193. [CrossRef]
3. Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4234–4245. [CrossRef] [PubMed]
4. Panagiotopoulos, D.; Orovas, C.; Syndoukas, D. Neural network based autonomous control of a speech synthesis system. *Intell. Syst. Appl.* **2022**, *14*, 200077. [CrossRef]

5. Bazzi, A.; Slock, D.T.; Meilhac, L. Sparse recovery using an iterative Variational Bayes algorithm and application to AoA estimation. In Proceedings of the 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, 12–14 December 2016; pp. 197–202. [CrossRef]
6. Bazzi, A.; Slock, D.T.M.; Meilhac, L. A Newton-type Forward Backward Greedy method for multi-snapshot compressed sensing. In Proceedings of the 2017 51st Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 29 October–1 November 2017; pp. 1178–1182. [CrossRef]
7. Li, N.; Liu, S.; Liu, Y.; Zhao, S.; Liu, M. Neural speech synthesis with transformer network. In Proceedings of the AAAI conference on artificial intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6706–6713. [CrossRef]
8. Lee, M.; Lee, J.; Chang, J.H. Non-autoregressive fully parallel deep convolutional neural speech synthesis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2022**, *30*, 1150–1159. [CrossRef]
9. Huang, R.; Zhao, Z.; Liu, H.; Liu, J.; Cui, C.; Ren, Y. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 2595–2605. [CrossRef]
10. Łańcucki, A. Fastpitch: Parallel text-to-speech with pitch prediction. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6588–6592. [CrossRef]
11. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), Sunnyvale, CA, USA, 13–15 September 2016; p. 125.
12. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
13. Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 8067–8077.
14. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech: Fast, Robust and Controllable Text to Speech. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
15. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010. [CrossRef]
16. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783. [CrossRef]
17. Liu, J.; Li, C.; Ren, Y.; Chen, F.; Zhao, Z. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11020–11028. [CrossRef]
18. Ren, Y.; Liu, J.; Zhao, Z. PortaSpeech: Portable and High-Quality Generative Text-to-Speech. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 13963–13974.
19. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 17022–17033.
20. Chen, N.; Zhang, Y.; Zen, H.; Weiss, R.J.; Norouzi, M.; Chan, W. WaveGrad: Estimating Gradients for Waveform Generation. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
21. Kim, H.; Kim, S.; Yoon, S. Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance. In Proceedings of the 39th International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 11119–11133.
22. Yang, X.; Zhou, D.; Feng, J.; Wang, X. Diffusion Probabilistic Model Made Slim. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 11–15 June 2023; pp. 22552–22562. [CrossRef]
23. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
24. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* **2023**, *56*, 1–51. [CrossRef]
25. Wang, W.; Bao, J.; Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; Li, H. SinDiffusion: Learning a Diffusion Model From a Single Natural Image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3412–3423. [CrossRef] [PubMed]
26. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.

27. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A Flow-based Generative Network for Speech Synthesis. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3617–3621. [CrossRef]
28. Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; et al. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 3918–3926.
29. Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; Courville, A.C. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
30. Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
31. Huang, R.; Lam, M.W.Y.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; Zhao, Z. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, Vienna, Austria, 23–29 July 2022; pp. 4157–4163. [CrossRef]
32. Rethage, D.; Pons, J.; Serra, X. A Wavenet for Speech Denoising. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5069–5073. [CrossRef]
33. Valentini-Botinhao, C.; King, S. Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 2746–2750. [CrossRef]
34. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
35. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
36. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; Volume 1, pp. 125–128. [CrossRef]
37. Kilgour, K.; Zuluaga, M.; Roblek, D.; Sharifi, M. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2350–2354. [CrossRef]
38. Choi, Y.; Jung, Y.; Suh, Y.; Kim, H. Learning to Maximize Speech Quality Directly Using MOS Prediction for Neural Text-to-Speech. *IEEE Access* **2022**, *10*, 52621–52629. [CrossRef]
39. Langlois, Q.; Jodogne, S. Practical Study of Deep Learning Models for Speech Synthesis. In Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, New York, NY, USA, 5–7 July 2023; pp. 700–706. [CrossRef]
40. Müller, M. Dynamic time warping. In *Information Retrieval for Music and Motion*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 69–84. [CrossRef]