



electronics

IMPACT
FACTOR
2.6

CITESCORE
6.1

Article

Fast Inference End-to-End Speech Synthesis with Style Diffusion

Hui Sun, Jiye Song and Yi Jiang

Special Issue

Deep Learning in Image Processing and Pattern Recognition, 2nd Edition

Edited by



Prof. Dr. Yuji Iwahori, Prof. Dr. Aili Wang, Prof. Dr. Haibin Wu and Dr. Xiaoming Sun



<https://doi.org/10.3390/electronics14142829>

Article

Fast Inference End-to-End Speech Synthesis with Style Diffusion

Hui Sun ¹, Jiye Song ^{2,*} and Yi Jiang ²¹ The Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentation of Heilongjiang Province, Harbin 150080, China; sunhui@hrbust.edu.cn² Department of Communications Engineering, Harbin University of Science and Technology, Harbin 150080, China; jasonj@hrbust.edu.cn

* Correspondence: 2320600072@stu.hrbust.edu.cn

Abstract

In recent years, deep learning-based end-to-end Text-To-Speech (TTS) models have made significant progress in enhancing speech naturalness and fluency. However, existing Variational Inference Text-to-Speech (VITS) models still face challenges such as insufficient pitch modeling, inadequate contextual dependency capture, and low inference efficiency in the decoder. To address these issues, this paper proposes an improved TTS framework named Q-VITS. Q-VITS incorporates Rotary Position Embedding (RoPE) into the text encoder to enhance long-sequence modeling, adopts a frame-level prior modeling strategy to optimize one-to-many mappings, and designs a style extractor based on a diffusion model for controllable style rendering. Additionally, the proposed decoder ConfoGAN integrates explicit F0 modeling, Pseudo-Quadrature Mirror Filter (PQMF) multi-band synthesis and Conformer structure. The experimental results demonstrate that Q-VITS outperforms the VITS in terms of speech quality, pitch accuracy, and inference efficiency in both subjective Mean Opinion Score (MOS) and objective Mel-Cepstral Distortion (MCD) and Root Mean Square Error (RMSE) evaluations on a single-speaker dataset, achieving performance close to ground-truth audio. These improvements provide an effective solution for efficient and controllable speech synthesis.

Keywords: text to speech; speech synthesis; end-to-end model; VITS; style diffusion

Academic Editor: Arkaitz Zubiaga

Received: 4 June 2025

Revised: 10 July 2025

Accepted: 11 July 2025

Published: 15 July 2025

Citation: Sun, H.; Song, J.; Jiang, Y. Fast Inference End-to-End Speech Synthesis with Style Diffusion. *Electronics* **2025**, *14*, 2829. <https://doi.org/10.3390/electronics14142829>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Speech synthesis, also known as Text-to-Speech (TTS), aims to generate natural and fluent speech waveforms from text. In recent years, deep neural network-based TTS models have significantly advanced the quality of synthesized speech, with end-to-end (E2E) models receiving increasing attention due to their streamlined architecture and strong joint optimization capabilities. Among them, Variational Inference Text-to-Speech (VITS) [1] has achieved high-quality and robust speech synthesis by integrating a Variational Autoencoder (VAE) [2], normalizing flow [3] and adversarial training [4] into a unified framework.

Despite the effectiveness of VITS in TTS tasks, several limitations remain:

1. The stochastic duration predictor performs poorly, often assigning excessively short durations to critical phonemes, which leads to noticeable phoneme skipping in the synthesized audio.
2. The relative positional encoding [5] used in the text encoder is computationally expensive and insufficient for modeling long-range dependencies, limiting the ability to capture natural prosody.

3. The GAN [4]-based decoder employs multiple transposed convolution layers, which are computationally expensive and form the primary bottleneck during training and inference.
4. The lack of explicit modeling of pitch information (e.g., fundamental frequency F_0) can lead to spectral artifacts and reduced synthesis quality.

To address these issues, Q-VITS, an improved end-to-end TTS model, is proposed with the following key enhancements:

1. Externally supervised duration modeling is employed to avoid irregular phoneme durations. Additionally, a frame-level prior modeling strategy is introduced by learning frame-wise mean and variance, which enhances the model's capacity to handle one-to-many mappings in acoustic modeling.
2. Rotary Position Embedding (RoPE) [6] is incorporated into the text encoder to strengthen long-sequence modeling capabilities and improve efficiency.
3. A diffusion-based Style Extractor is integrated to capture speaker style and expressive characteristics with high fidelity, significantly improving the expressiveness and controllability of synthesized speech.
4. A novel decoder architecture, ConfoGAN, is designed to overcome the efficiency and modeling limitations of the original decoder. Built upon BigVGAN [7], ConfoGAN integrates the several following critical components.

An advanced transformer module, Conformer [8], to model long-term temporal dependencies in speech.

A Source Generator module based on the Harmonic-plus-Noise Source-Filter (Hn-NSF) [9] model is employed to explicitly incorporate fundamental frequency (F_0) information.

The inverse short-time Fourier transform (iSTFT) [10] and multi-band processing mechanism using Pseudo-Quadrature Mirror Filter (PQMF) [11] is employed to decompose the waveform into multiple sub-bands for parallel generation, thereby significantly improving inference efficiency.

The experimental results show that Q-VITS significantly outperforms the original VITS in terms of speech naturalness, pitch accuracy, voice feature similarity, and generation efficiency, validating the effectiveness of the proposed method.

2. Related Work

Over the past few decades, Concatenative Speech Synthesis [12,13] and Statistical Parametric Speech Synthesis (SPSS) [14,15] have been the two dominant paradigms in speech synthesis. Traditional approaches, such as SPSS, are typically composed of multiple independent modules, including text analysis and linguistic feature extraction (e.g., duration and fundamental frequency prediction), acoustic modeling (generating acoustic representations from linguistic inputs), and a vocoder (converting acoustic features into waveforms). Although these methods are well-established from an engineering perspective, the loose coupling between modules and the inconsistency of training objectives often result in synthesized speech that lacks naturalness and expressiveness.

In recent years, the emergence of End-to-End (E2E) TTS models has significantly advanced the development of speech synthesis technologies. These models can directly generate speech waveforms or intermediate spectral representations (e.g., Mel spectrograms [16]) from raw text inputs such as characters or phonemes. This paradigm effectively simplifies the overall system architecture, reduces reliance on expert knowledge and manual annotations, and improves both training efficiency and speech naturalness. For example, Char2Wav [17] developed a model that integrates attention-driven encoder-decoder learning for end-to-end speech generation from text. Tacotron [18] was the first to predict linear

spectrograms directly from textual input, while Tacotron2 [19] further improved speech naturalness, yielding results that closely resemble human speech.

With increasing emphasis on generation efficiency, Non-Autoregressive (NAR) TTS models have gained prominence. By removing the sequential dependency inherent in autoregressive approaches, these models enable parallel generation, significantly improving inference speed and mitigating the problem of error accumulation. For instance, FastSpeech [20] adopts a Transformer-based parallel architecture [21] to generate Mel-spectrograms, and FastPitch [22] introduces explicit pitch prediction to enhance prosodic expressiveness. Building upon this, FastSpeech2 [23] incorporates additional auxiliary features such as duration, pitch, and energy as conditional inputs, substantially improving both controllability and synthesis quality.

To further improve the accuracy of acoustic modeling, recent studies have explored incorporating more prosodic information on the text side. The objective is to bridge the representation gap between textual input and acoustic output, thereby reducing modeling complexity. For example, both FastPitch and FastSpeech2 demonstrate that incorporating prosodic cues like pitch (F_0) and energy leads to noticeable improvements in speech naturalness.

Despite the promising performance of these approaches, certain limitations persist. Most systems still adopt a two-stage architecture, where an acoustic model first generates intermediate representations and then a vocoder converts them into waveforms. While this modular design offers flexibility, the decoupled training process and the reliance on hand-crafted intermediate representations constrain the model's ability to capture fine-grained speech details.

To overcome these limitations, VITS [1] introduces a single-stage, fully parallel, end-to-end speech synthesis framework. By jointly training a Variational Autoencoder (VAE), a normalizing flow, and a Generative Adversarial Network (GAN) [4], VITS achieves a unified optimization of acoustic modeling and waveform generation. This integration leads to notable improvements in both speech naturalness and model robustness. Experimental results show that VITS outperforms traditional two-stage models on multiple public benchmarks, marking a significant advancement in end-to-end speech synthesis.

Rotary position embedding (RoPE) has been widely applied in various Transformer [21]-based models such as ChatGLM [24] and LLaMA [25]. Matcha-TTS [26] introduces RoPE into diffusion probabilistic models for TTS. Compared to conventional relative position encoding, RoPE is more efficient in both computation and memory consumption, and it generalizes better to longer sequences.

Diffusion models have demonstrated strong potential in Text-to-Speech (TTS) tasks due to their high-quality voice style transfer, diverse generative capabilities, and shorter training time. However, slow inference speed, high computational cost, and limited controllability over voice style and speaker characteristics remain significant challenges to be addressed. To this end, StyleTTS2 [27] proposes a hybrid approach that samples a fixed-length style vector through the diffusion process and combines it with a GAN-based speech synthesizer, thereby significantly improving synthesis efficiency and expressiveness.

Graph neural networks have shown promising results in modeling complex speech features beyond traditional sequential methods. Li et al. [28] proposed a Graph-LSTM architecture for speech emotion recognition, leveraging frame-level graph construction and LSTM-based aggregation with weighted pooling to effectively capture temporal dependencies and emotional cues. This approach inspires our design of the style encoder for capturing expressive speech characteristics.

This study is inspired by the aforementioned works but introduces two fundamental innovations.

First, we incorporate diffusion-based style modeling into the core end-to-end TTS framework to provide style conditioning, enabling high-fidelity expressive control with minimal impact on generation efficiency.

Second, we propose a novel decoder architecture, ConfoGAN, which combines BigVGAN, Conformer modules, and a source generator based on the Harmonic-plus-Noise Source-Filter (Hn-NSF) model to achieve efficient and high-quality waveform synthesis.

3. Materials and Methods

The proposed model, named Q-VITS, adopts a Conditional Variational Autoencoder (CVAE) [29] framework for end-to-end speech synthesis. It consists of three main components: prior encoder, posterior encoder, and decoder. The training procedure is illustrated in Figure 1, and the inference procedure is shown in Figure 2. The posterior encoder extracts the latent representation z from the linear spectrogram, while the decoder reconstructs the speech waveform from z . The prior encoder provides a prior distribution for z conditioned on the input text. The architectural details of the prior encoder, posterior encoder and decoder are elaborated in the subsequent subsections.

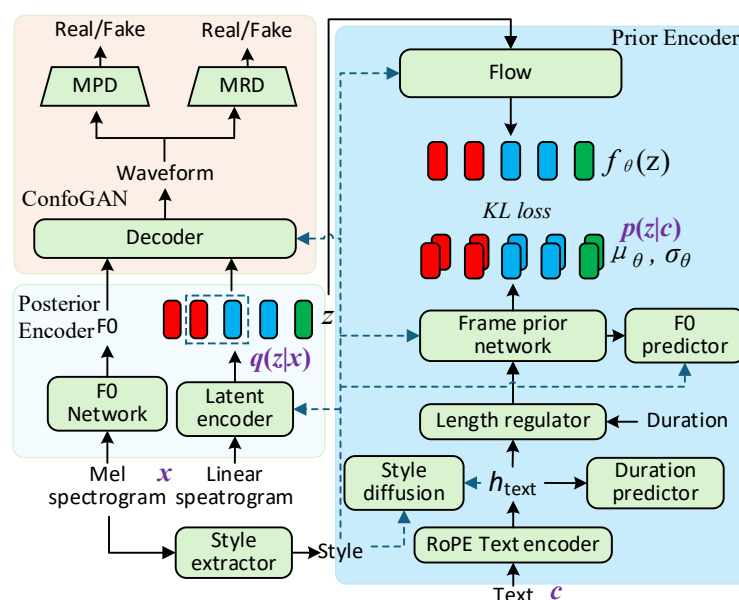


Figure 1. Training procedure.

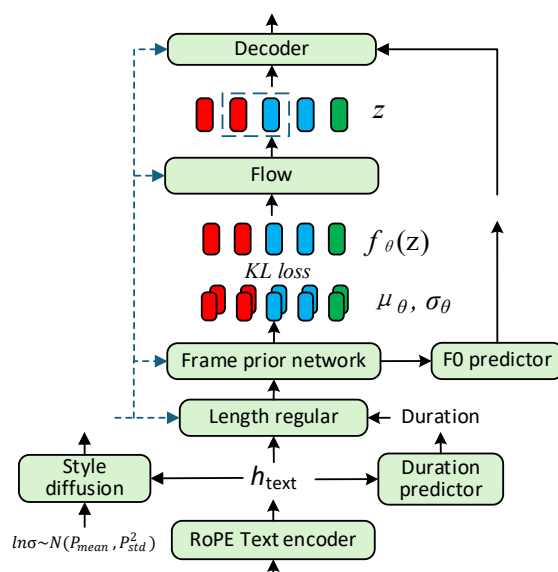


Figure 2. Inference procedure.

3.1. Prior Encoder

The prior encoder predicts the prior distribution $p(z|c)$ from the phoneme sequence c , serving as the prior regularization term in the Conditional Variational Autoencoder (CVAE) [29]. Due to the significant acoustic variability in speech, different frames within a single phoneme may follow different distributions. Inspired by VISinger [30], a frame prior network (FPN) is introduced to model frame-level mean and variance.

The use of the FPN requires knowledge of the duration associated with each phoneme. Therefore, the duration predictor must be trained using annotated duration labels, making it incompatible with the unsupervised alignment approach used in VITS. To address this, Q-VITS adopts Montreal Forced Alignment (MFA) [31] to obtain annotated labels, which provide the duration information necessary for alignment during training.

The decoder of Q-VITS is required to model the fundamental frequency F_0 . However, since no reference audio is available during inference, F_0 must be predicted from the text. To achieve this, an F_0 predictor is incorporated, which utilizes hidden states extracted from the frame prior network to estimate fundamental frequency at the frame level. During inference, the predicted F_0 is used as an input to the decoder described in Section 3.3.

RoPE Text Encoder

The text encoder serves as the first module in the prior encoder, as illustrated in Figure 1. It takes the preprocessed text sequence as input and generates phoneme-level representations h_{text} based on the Transformer architecture. Each phoneme is encoded as a 192-dimensional vector. Inspired by RoFormer [6], our model incorporates enhancements to the original Transformer architecture, as shown in Figure 3. Specifically, it replaces the relative positional encoding in the attention layers with Rotary Position Embedding (RoPE) to improve the modeling capability for long sequences.

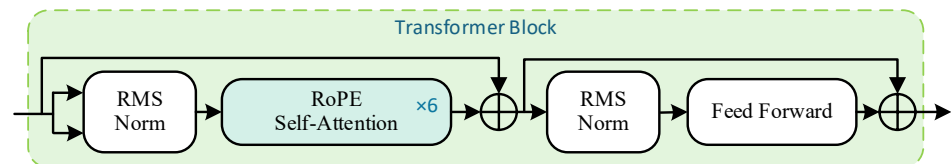


Figure 3. Transformer block structure.

3.2. Posterior Encoder

As illustrated in Figure 1, the posterior encoder consists of two components. The first is the Latent Encoder, which estimates the posterior distribution $q(z|x)$ from the linear spectrogram. A latent variable z is sampled from this distribution and aligned with the prior distribution $p(z|c)$ derived from the text during training. The second component is the F_0 Network, which extracts the fundamental frequency F_0 from the Mel spectrogram. The obtained z and F_0 are jointly fed into the decoder to reconstruct the original speech waveform. This architecture enables effective optimization of the GAN-based decoder during training, thereby enhancing the perceptual quality of generated audio.

3.3. Decoder

In VITS, the decoder adopts the generator architecture from HiFi-GAN [32]. This decoder progressively upsamples latent features z through multiple layers of transposed convolution to generate time-domain signals at the same sampling rate as the target speech waveform. As the final component in a speech synthesis system, the decoder plays a crucial role in determining the naturalness, clarity, and temporal consistency of the generated audio.

Despite its effectiveness in improving speech quality, the decoder in VITS exhibits the following four major limitations, as observed in preliminary experiments:

1. **Artifact Generation.** The use of transposed convolution during upsampling tends to introduce high-frequency noise and pitch-related artifacts, which significantly degrade the clarity and naturalness of the synthesized speech [33].
2. **Low Inference Efficiency.** The complex convolutional architecture incurs substantial computational overhead. Experiments show that this module accounts for over 96% of the total inference time [34], becoming the primary bottleneck of the system.
3. **Limited F_0 Modeling Capability.** Downsampling mechanisms such as average pooling and uniform sampling often lead to aliasing, making it difficult to accurately reconstruct the fundamental frequency (F_0) and its harmonic structure [35]. This issue is especially pronounced in speech with significant pitch variations [36,37].
4. **Inadequate Long-Term Dependency Modeling.** Although GAN-based decoders are effective at generating fine-grained details, their reliance on convolutional structures with limited receptive fields hinders their ability to capture long-term patterns.

To address these issues, a novel neural decoder architecture named ConfoGAN is proposed as the core decoder module of Q-VITS. Built upon the BigVGAN [7] framework, ConfoGAN integrates five key techniques: periodic source modeling based on Neural Source-Filter (NSF), Anti-Aliasing Activation functions, learnable multi-band synthesis, and long-term dependency modeling via Conformer with Ring Attention [6]. This design achieves substantial improvements in both speech quality and inference efficiency. The structure of the decoder is shown in Figure 4.

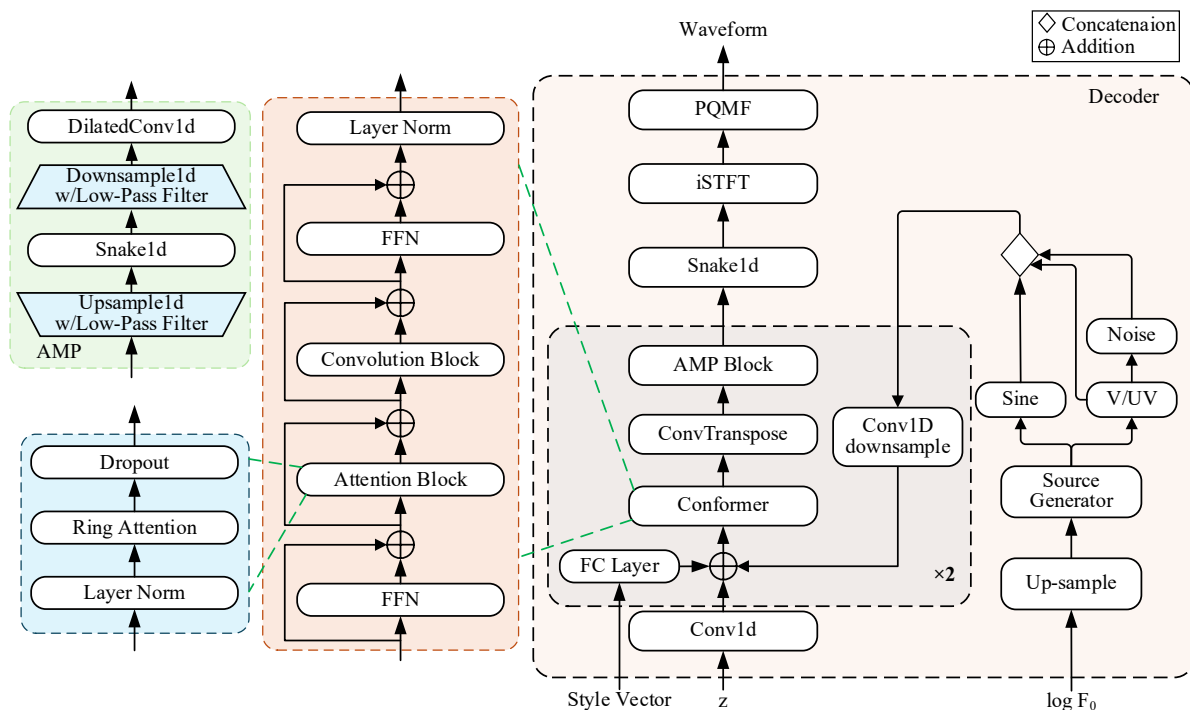


Figure 4. ConfoGAN structure.

Firstly, the Leaky ReLU activation function used in HiFi-GAN lacks the capability to model periodic structures in speech, such as fundamental frequency (F_0) and harmonics. This limitation often results in artifacts in the generated waveforms. To address this issue, we adopt the Snake activation function [38], as introduced in BigVGAN, which is defined as follows:

$$f_{\alpha}(x) = x + \frac{1}{\alpha} \sin^2(\alpha x) \quad (1)$$

which incorporates a learnable frequency parameter to introduce periodic components. This enhances the model's ability to represent multi-frequency periodic structures in speech.

However, applying the Snake activation directly can generate high-frequency components beyond the discrete sampling capacity, causing aliasing artifacts. To mitigate this, we implement the anti-aliasing strategy from BigVGAN, which upsamples the input signal by a factor of two along the temporal axis, applies the Snake activation combined with low-pass filter, and then downsamples back to the original rate.

In ConfoGAN, all residual dilated convolution layers adopt an anti-aliasing periodic activation function, which is modularized into the Anti-Aliased Multi-Period (AMP) module. As shown in Figure 4, the AMP block is composed of multiple AMP sub-modules connected via residual connections, each employing periodic activation functions with different frequencies to model periodic patterns across various frequency ranges.

To explicitly incorporate fundamental frequency (F_0) information, the design of HiFT-Net [39] is adopted, and a Source Generator module based on the Neural Harmonic-plus-Noise Source Filter (Hn-NSF) [9] is integrated at each upsampling stage. This module takes the upsampled F_0 contour as input and generates the corresponding harmonic sinusoidal excitation signal (*Sine*) along with a voiced/unvoiced (v/uv) flag. Subsequently, the *Sine*, v/uv , and a noise signal randomly sampled from the *Sine* are concatenated to form a continuous excitation source. Finally, the excitation source is mapped through a convolutional downsampling layer and fused with the main branch features z .

To accelerate inference, the last two layers of transposed convolutions in the original decoder are replaced with Inverse Short-Time Fourier Transform (iSTFT) [10] and Learnable Pseudo-Quadrature Mirror Filter (Learnable PQMF) [11]. The iSTFT restores the magnitude and phase information of the model-generated spectrogram into sub-band time-domain waveforms, and the Learnable PQMF structure is used to synthesize multi-band waveforms. Experiments show that this multi-band reconstruction strategy improves inference speed by approximately $4.4\times$ while maintaining the naturalness of synthesized speech.

A sub-band multi-resolution STFT loss is introduced to provide finer spectral supervision across multiple time-frequency resolutions during training, thereby improving the reconstruction quality of VITS.

To address the challenge of modeling long-term dependencies, an enhanced Transformer structure, named Conformer, is introduced before the transposed convolution layers in the decoder. Conformer combines the local pattern modeling strengths of convolutional networks with the global sequence modeling capabilities of self-attention mechanisms, achieving a better balance in temporal modeling. The Ring Attention [6] mechanism within Conformer is a highly efficient form of local attention that performs attention computations within a restricted local context. This significantly reduces computational cost while retaining awareness of global information, allowing the model to model long-term contextual relationships without sacrificing efficiency, and thus compensating for the limitations of GAN-based vocoders in this aspect.

Overall, the ConfoGAN decoder, by incorporating five key mechanisms—fundamental frequency modeling, anti-aliasing activation, multi-band reconstruction, frequency-domain synthesis, and long-term dependency modeling—significantly improves both speech quality and generation speed, demonstrating its practicality in high-fidelity TTS systems.

3.4. Style Modeling

To enhance the capability of style imitation in speech synthesis, Q-VITS introduces the novel combination of a Style Extractor and a Style Diffusion Mechanism.

While prior works commonly employ Variational Autoencoder (VAE) or normalizing flows for style modeling, these approaches present notable limitations. VAEs often suffer

from posterior collapse [40], resulting in reduced diversity and overly averaged styles. Normalizing flows require invertible network architectures, leading to increased model complexity, slower inference speed, and unstable training.

In contrast, diffusion models inherently provide sampling diversity and training stability, enabling more natural modeling of non-linguistic information such as prosodic variations and emotional fluctuations in speech. By progressively refining noisy latent variables, the diffusion-based style module in Q-VITS captures richer expressive details and achieves better trade-offs between style flexibility and speech quality, while maintaining compatibility with components like pitch prediction and frame-level synthesis.

The Style Extractor models the speaker's style as a latent random variable. The Style Diffusion component performs conditional sampling based on a probabilistic diffusion model [41]. The architectures of the Style Extractor and Style Diffusion are illustrated in Figures 5 and 6, respectively.

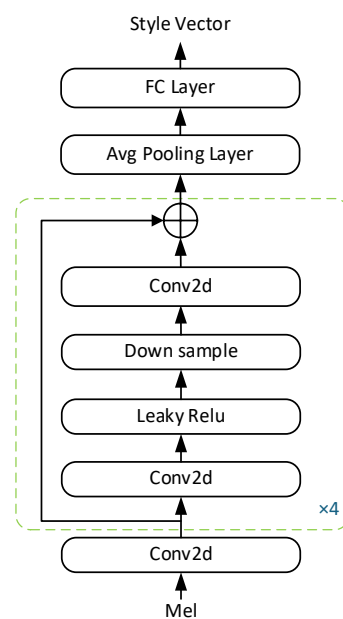


Figure 5. Style Extractor structure.

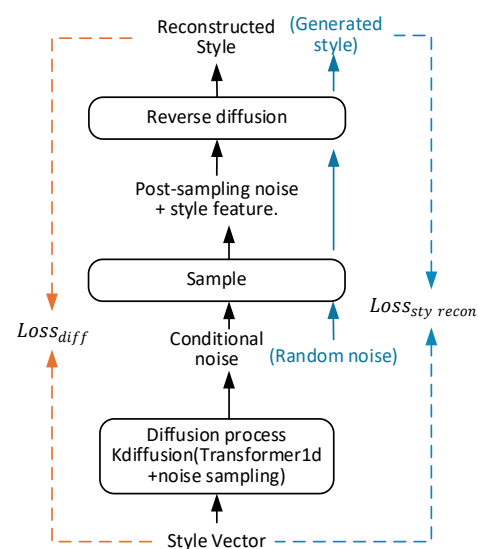


Figure 6. Style Diffusion process.

The Style Extractor adopts a multi-layer convolutional architecture equipped with spectral normalization and Leaky ReLU activation. It extracts features from randomly

cropped mel-spectrogram segments to produce a fixed-dimensional vector that captures speaker timbre and intonation. This vector serves both as conditional information for subsequent modules and as input to the Style Diffusion module.

The style diffuser generates dynamic style vectors based on the random noise from input. During training, the model is jointly optimized through a forward noise injection and reverse denoising process, guided by two objectives: the Style Diffusion loss and the style reconstruction loss. The Style Diffusion loss L_{diff} measures the discrepancy between the predicted style vector and the target style, while the style reconstruction loss $L_{sty recon}$ constrains the reconstruction error between the style vector generated from noise and the reference style. The formulations of these two loss functions are as follows:

$$L_{diff} = \|\widehat{cond} - cond\|_1 \quad (2)$$

$$L_{sty recon} = \|cond - recon\|_1 \quad (3)$$

During inference, the Style Diffusion module conditions the input text and progressively reconstructs the target style vector from random noise through multi-step sampling. This enables the synthesized speech to dynamically adjust prosodic attributes such as timbre and intonation according to the context, achieving more natural and flexible voice style control.

Compared to traditional approaches, this mechanism leverages the synergy between encoding and diffusion to enhance the realism of pronunciation details while maintaining style consistency across utterances. The progressive generation nature of the diffusion model improves the modeling of complex acoustic patterns, while the conditional generation framework ensures compatibility with other components such as pitch prediction and frame-level synthesis.

3.5. Final Loss

With the above CVAE and adversarial training, the proposed model is optimized with the full objective:

$$L_{total} = L_{recon} + L_{kl} + L_{dur} + L_{adv(G)} + L_{fm(G)} + L_{pitch} + L_{stft} + L_{diff} + L_{sty recon} \quad (4)$$

L_{kl} is the KL divergence between the prior z and the posterior z .

L_{dur} is the L2 duration loss, supervising the duration model.

$L_{adv(G)}$ is the adversarial loss from GAN, encouraging the generation of realistic waveforms.

$L_{fm(G)}$ is the feature matching loss, which promotes similarity between generated and real data in feature space, improving training stability.

L_{pitch} is the pitch accuracy loss, which ensures the generated waveform maintains accurate pitch characteristics.

4. Experiments

4.1. Datasets

In this study, we employed two publicly available datasets for model training and evaluation: LJSpeech and Databaker. The LJSpeech dataset was recorded by a female speaker. It contains 13,100 audio clips with a total duration of 25 h approximately [42]. The sample rate is 22.05 kHz and quantized to 16 bit.

The Databaker dataset [43] is a high-quality Mandarin speech corpus provided by Beijing Databaker Technology Co., Ltd. It contains approximately 12 h of Chinese speech

recorded by a single female speaker. The recordings span various daily scenarios, such as news broadcasts and movie dialogs. All audio samples were resampled to 24 kHz.

In the experiments, 5% of the samples in each dataset were randomly selected as the validation set, and the remaining samples were used for training.

The main reason for choosing LJSpeech and Databaker is that the core objective of this study is to evaluate the effectiveness of the proposed multi-component model in improving speech synthesis quality and style control, rather than optimizing for multi-speaker tasks. Using single-speaker datasets helps eliminate variability caused by different speakers, allowing for a more focused assessment of the model's improvements in audio quality, prosody, and expressiveness. Furthermore, both datasets are widely used public benchmarks in the TTS field, enabling fair comparisons with existing methods.

4.2. Model Setups

To validate the effectiveness of each improvement in Q-VITS, several models are prepared for comparison. Considering that BigVGAN outperforms HiFi-GAN in speech generation quality, we decided to conduct experiments using VITS with a BigVGAN decoder. The configurations of the models are shown in Table 1. In the table,

VITS refers to the baseline model, using the BigVGAN decoder;

FPN-VITS introduces a frame prior network based on VITS;

RoP-VITS applies Rotary Positional Encoding (RoPE) in the text encoder of FPN-VITS;

F0-VITS incorporates fundamental frequency (F_0) information into RoP-VITS;

C-VITS replaces the original decoder in F0-VITS with the proposed ConfoGAN structure; Q-VITS further enhances C-VITS by introducing a Style Extractor and Style Diffusion Mechanism.

Table 1. Configurations of the models. FPN: frame prior network in VISinger [30]. RoPE: Rotary Positional Embedding. Style: Style Extractor and Style Diffusion.

Model	FPN	RoPE	F0 input	ConfoGAN	Style
VITS	×	×	×	×	×
FPN-VITS	✓	×	×	×	×
RoP-VITS	✓	✓	×	×	×
F0-VITS	✓	✓	✓	×	×
C-VITS	✓	✓	✓	✓	×
Q-VITS	✓	✓	✓	✓	✓

Different models are associated with specific loss functions and corresponding loss coefficients, as shown in Table 2.

Table 2. Configuration of loss functions and their corresponding loss coefficients across models.

Model	L_{recon}	L_{kl}	L_{dur}	$L_{adv(G)}$	$L_{fm(G)}$	L_{pitch}	L_{stft}	L_{diff}	$L_{sty recon}$
VITS	45	1	1	1	2	-	-	-	-
FPN-VITS	45	1	1	1	2	-	-	-	-
RoP-VITS	45	1	1	1	2	-	-	-	-
F0-VITS	45	1	1	1	2	1	-	-	-
C-VITS	45	1	1	1	2	1	1	-	-
Q-VITS	45	1	1	1	2	1	1	1	1

The decoder employs a two-layer upsampling process with scaling factors [3,5] and kernel sizes [6,10], progressively increasing the temporal resolution of the audio. The Conformer module is applied in the first upsampling layer. The sub-band decoder divides the full-band audio into four sub-bands to improve speech reconstruction quality.

The linear spectrogram, used as the input to the posterior encoder, is computed via STFT with an FFT size of 1024, window size of 960, and hop size of 240.

Mixed precision training is employed during the training process. The batch size is set to 4, and all models are trained for 100,000 steps.

5. Results

5.1. Evaluation

To facilitate a comprehensive performance comparison, speech samples were generated using 100 text inputs from the test set. The evaluation criteria include both subjective and objective metrics.

Subjective evaluation was conducted using a 5-point Mean Opinion Score (MOS) [44]. After listening to each synthesized utterance, participants rated its naturalness on a scale from 1 (poor) to 5 (excellent), with 0.5-point intervals. Each sample was rated independently by multiple listeners, and the average score was reported as the MOS for that utterance, along with the 95% confidence interval.

Objective evaluation includes Mel-Cepstral Distortion (MCD) [45] and logF0 Root Mean Square Error (logF0 RMSE) to assess speech quality.

MCD measures the average distance between the Mel-Frequency Cepstral Coefficients (MFCCs) of the synthesized and ground truth speech, with smaller values indicating higher spectral similarity.

logF0 RMSE evaluates the deviation between the predicted and reference fundamental frequency (F0) values, reflecting pitch accuracy. Lower logF0 RMSE indicates better prosodic consistency and more natural intonation in the synthesized speech.

To evaluate inference efficiency, we measured the Real-Time Factor (RTF) [46] of five models using test sentences on a CPU. RTF is defined as the ratio of the time required to synthesize an utterance to the actual duration of that utterance. The average RTF of over 100 utterances is reported. The results are presented in Tables 3 and 4.

Table 3. The evaluation results of models on the LJSpeech dataset, the best scores emphasized using boldface.

Model	MOS \pm CI	MCD	logF0 RMSE	RTF
Ground truth	4.32	-	-	-
VITS	3.89	6.89	0.2889	0.223
FPN-VITS	3.91	6.88	0.2362	0.221
RoP-VITS	3.93	6.73	0.2274	0.216
F0-VITS	4.02	6.71	0.2203	0.276
C-VITS	4.08	6.62	0.2194	0.063
Q-VITS	4.12	6.36	0.2142	0.072

Table 4. The evaluation results of models on the Databaker dataset, the best scores emphasized using boldface.

Model	MOS \pm CI	MCD	logF0 RMSE	RTF
Ground truth	4.33	-	-	-
VITS	3.73	8.15	0.2803	0.228
FPN-VITS	3.77	8.07	0.2367	0.217
RoP-VITS	3.82	7.85	0.2342	0.211
F0-VITS	3.89	7.66	0.2309	0.222
C-VITS	4.01	7.59	0.2282	0.061
Q-VITS	4.09	7.48	0.2221	0.070

As shown in Tables 3 and 4, the speech generated by Q-VITS performs excellently in all evaluation metrics, with the only exception being inference speed (RTF). The comparison between F0-VITS and C-VITS shows that C-VITS significantly reduces inference time by using a ConfoGAN-based decoder (77.2% reduction on the LJSpeech dataset and 72.5% reduction on the Databaker dataset), which validates the effectiveness of this method in improving inference speed. Meanwhile, the improvements in MOS and MCD indicate that the speech quality has been optimized.

Compared to C-VITS, Q-VITS requires diffusion sampling to obtain style vectors and add them to subsequent synthesis modules, which increases inference time. However, Q-VITS still shows significantly lower inference time than VITS, FPN-VITS, and RoP-VITS (67.7%, 67.4%, and 66.7% reduction in the LJSpeech dataset, and 69.3%, 67.7%, and 66.8% reduction in the Databaker dataset), demonstrating its advantage in inference efficiency.

The comparison of MOS, MCD, and $\log F_0$ RMSE between FPN-VITS and VITS shows that introducing the frame prior network helps to improve speech quality, particularly in fundamental frequency modeling, with significant improvements.

The comparison between C-VITS and Q-VITS shows that adding additional style references helps generate more natural speech, making the synthesized speech's speaker characteristics closer to the target speaker, further enhancing the naturalness and authenticity of the speech.

Table 5 presents the inference parameters and inference computational costs (GFLOPs) of various VITS variants, along with the percentage differences relative to the original VITS model for each metric, allowing for an intuitive comparison.

Table 5. Inference parameters with GFLOPs for VITS variants.

Model	Inference Params	GFLOPs
VITS	28.26 M	59.14
FPN-VITS	24.82 M	53.74
RoP-VITS	19.51 M	49.69
F0-VITS	21.24 M	51.34
C-VITS	23.93 M	13.43
Q-VITS	28.89 M	15.14

Inference parameters refer to the actual number of parameters used during inference, reflecting the resource consumption when deploying the model.

Inference computational cost is measured in GFLOPs, indicating the number of floating-point operations required for a single forward pass and reflecting the model's inference efficiency.

The inference parameter count of C-VITS is 23.93 M, slightly higher than that of the original F0-VITS (21.24 M), representing an increase of approximately 12.7%. However, the inference computational cost of C-VITS is only 13.43 GFLOPs, a reduction of 73.8% compared to F0-VITS (51.34 GFLOPs), demonstrating a significant advantage in computational efficiency. This indicates that the use of inverse short-time Fourier transform and multi-band audio synthesis in the ConfoGAN-based decoder of C-VITS plays a crucial role in reducing computational costs.

The inference parameter count of Q-VITS is 21% higher than that of C-VITS, while its inference computational cost is only 12.7% higher, showing a relatively modest increase. This is because Q-VITS, during inference, only utilizes the sampler of the Style Diffusion module to sample from the latent random variable, generating the style vector used for waveform reconstruction, thus effectively controlling computational costs.

The introduction of Rotary Positional Embedding (RoPE) significantly reduces both the model's parameter count and computational cost, with reductions of approximately 21.4% and 7.5%, respectively. Both RoP-VITS and F0-VITS exhibit lower inference parameters and computational costs than the original VITS, accounting for approximately 69–75% and 11–14% of the original model, respectively.

The comparison of speech synthesis quality across different models is shown in Table 6. The models compared include Tacotron, Transformer, and FastSpeech. As shown in Table 6, the proposed method outperforms other speech synthesis methods in terms of synthesized speech quality and achieves a good inference speed.

Table 6. Comparison of metrics across different models, the best scores emphasized using boldface.

Model	Baseline	MOS	MCD	RTF
Tacotron 2	Tacotron	3.77	7.17	0.33
Transformer TTS	Transformer	3.74	7.43	0.826
FastSpeech	Transformer TTS	3.96	6.96	0.048
FastSpeech 2	FastSpeech	4.11	6.83	0.021
FastPitch	FastSpeech	4.06	6.87	0.029
Q-VITS	VITS	4.12	6.36	0.072

Mel spectrograms have been widely used in audio deep learning tasks and are a better audio feature representation than the others. This study also conducted a visual comparison of the activation maps between the models. Figure 7 presents the Mel spectrograms of speech waveforms synthesized by the RoP-VITS and the F0-VITS models.

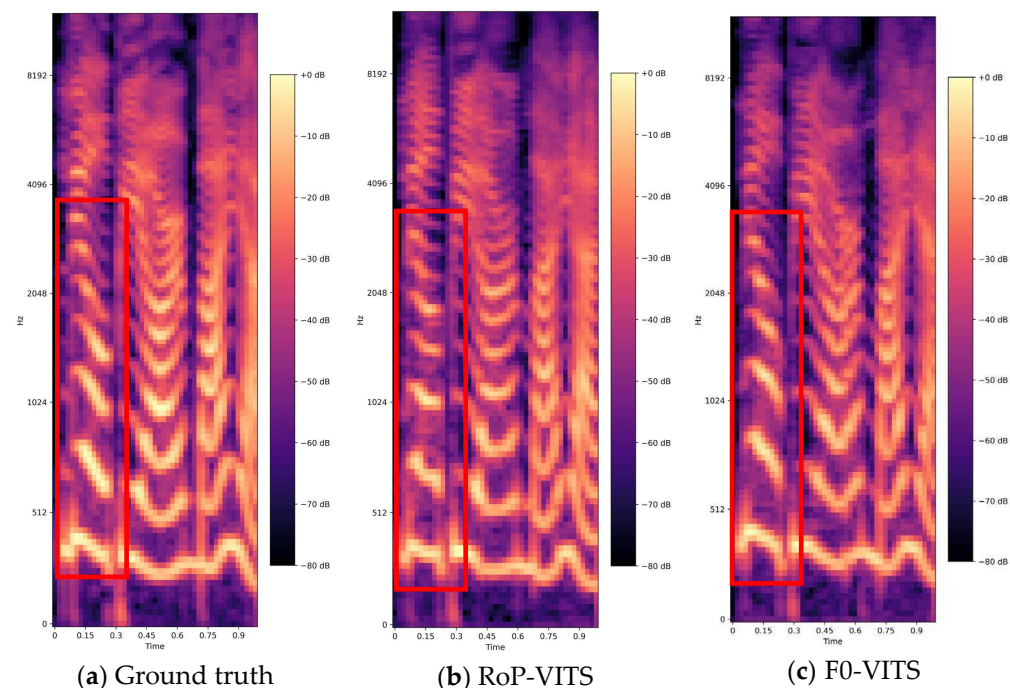


Figure 7. Mel-spec of the synthesized and ground truth.

As highlighted by the red boxes in the figure, the RoP-VITS model exhibits poor pitch fitting with an unstable pitch contour. In contrast, the Mel-spectrogram produced by F0-VITS appears smoother in the corresponding regions, which is consistent with the trend in the ground truth. As shown in Table 3, the MCD and logF0 RMSE of F0-VITS are 6.71 and 0.2209, respectively, both better than those of RoP-VITS, which are 6.73 and 0.2274.

This phenomenon is mainly caused by the fact that RoP-VITS does not feed the F_0 predicted by the prior encoder into the decoder, resulting in an unstable pitch contour in the synthesized speech.

5.2. Discussion

One limitation of the current study is that the evaluation is conducted on single-speaker datasets (LJSpeech and Databaker), which restrict the assessment to relatively homogeneous speech scenarios. Consequently, the proposed model's ability to generalize to cross-speaker synthesis, different genders, and various accents has not been fully explored. This limitation may introduce some uncertainty and constrain the applicability of the current conclusions when extended to multi-speaker or more diverse real-world settings.

To address this issue, future work will focus on extending and evaluating the proposed model on multi-speaker datasets such as VCTK and LibriTTS. This will help further validate the model's generalizability and enhance its adaptability and synthesis quality in cross-speaker and cross-style generation tasks.

6. Conclusions

Building upon the robust modeling capabilities of VITS, Q-VITS introduces several innovative modules that notably enhance speech naturalness, style expressiveness, and generation efficiency. By combining RoPE for positional encoding, F_0 prediction, diffusion-based style modeling, and the Conformer-powered GAN decoder-ConfoGAN, Q-VITS excels at capturing prosody and expressive nuances while effectively avoiding the high-frequency artifacts and instability often seen in conventional decoders. Particularly, the inclusion of ConfoGAN alleviates computational bottlenecks and boosts inference speed without sacrificing audio quality.

Beyond its effectiveness in TTS, ConfoGAN's modular and efficient design offers promising potential for generalization to other generative tasks, such as singing voice synthesis, emotional voice conversion, and audio-driven animation. This universality highlights its adaptability to a broader range of sequence-to-sequence generation problems.

For future work, we plan to enhance Q-VITS in two directions:

Firstly, by incorporating speaker embedding into the style encoder, the model can be extended to support explicit speaker control and zero-shot speaker adaptation.

Secondly, by adapting the system to multilingual data, through the integration of language embeddings or language-adaptive pre-training techniques, Q-VITS could be further extended to handle diverse linguistic inputs, enabling more versatile and inclusive speech synthesis across languages.

Overall, Q-VITS offers a practical, high-fidelity, and extensible architecture for next-generation TTS systems.

Author Contributions: Methodology, H.S.; software, J.S.; writing—original draft, J.S. and Y.J.; visualization, J.S.; supervision, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: Supported by the 2022 New Round of Heilongjiang Province “Double First-Class” Discipline Collaborative Innovation Achievement Projects (Grant No. LJGXCG2022-067) and the Harbin Manufacturing Industry Science and Technology Innovation Talents Project (Grant No. 2023CXRC021).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://keithito.com/LJ-Speech-Dataset/> and <https://en.data-baker.com/datasets/freeDatasets/>, accessed on 3 June 2025.

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

References

- Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the International Conference on Machine Learning, PMLR, Beijing, China, 18–24 July 2021; pp. 5530–5540.
- Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
- Rezende, D.; Mohamed, S. Variational inference with normalizing flows. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 1530–1538.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680. [[CrossRef](#)]
- Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv* **2018**, arXiv:1803.02155.
- Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [[CrossRef](#)]
- Lee, S.; Ping, W.; Ginsburg, B.; Catanzaro, B.; Yoon, S. Bigvgan: A universal neural vocoder with large-scale training. *arXiv* **2022**, arXiv:2206.04658.
- Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
- Wang, X.; Takaki, S.; Yamagishi, J. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *28*, 402–415. [[CrossRef](#)]
- Kaneko, T.; Tanaka, K.; Kameoka, H.; Seki, S. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6207–6211.
- Nguyen, T.Q. Near-perfect-reconstruction pseudo-QMF banks. *IEEE Trans. Signal Process.* **1994**, *42*, 65–76. [[CrossRef](#)]
- Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, GA, USA, 9 May 1996; IEEE: Piscataway, NJ, USA, 1996; pp. 373–376.
- Black, A.W.; Taylor, P. Automatically clustering similar units for unit selection in speech synthesis. In Proceedings of the EUROSPEECH'97: 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997.
- Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [[CrossRef](#)]
- Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 7962–7966.
- Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [[CrossRef](#)]
- Saito, Y.; Takamichi, S.; Saruwatari, H. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *26*, 84–96. [[CrossRef](#)]
- Wang, Y.; Skerry-Ryan, R.J.; 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. *arXiv* **2017**, arXiv:1703.10135.
- Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z. 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; IEEE: Piscataway, NJ, USA, 2018.
- Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. FastSpeech: Fast, robust and controllable text to speech. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3165–3174.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
- 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, 13 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6588–6592.
- 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. *arXiv* **2020**, arXiv:2006.04558.
- GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv* **2024**, arXiv:2406.12793.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
- Mehta, S.; Tu, R.; Beskow, J.; Székely, É.; Henter, G.E. Matcha-TTS: A fast TTS architecture with conditional flow matching. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 11341–11345.

27. Li, Y.A.; Han, C.; Raghavan, V.; Mischler, G.; Mesgarani, N. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 19594–19621.
28. Li, Y.; Wang, Y.; Yang, X.; Im, S.K. Speech emotion recognition based on Graph-LSTM neural network. *EURASIP J. Audio Speech Music Process.* **2023**, *2023*, 40. [[CrossRef](#)]
29. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Systems* **2015**, *2*, 3483–3491.
30. Zhang, Y.; Cong, J.; Xue, H.; Xie, L.; Zhu, P.; Bi, M. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7237–7241.
31. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kald. *Interspeech* **2017**, *2017*, 498–502.
32. Kong, J.; Kim, J.; Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17022–17033.
33. Pons, J.; Pascual, S.; Cengarle, G.; Serrà, J. Upsampling artifacts in neural audio synthesis. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3005–3009.
34. Kawamura, M.; Shirahata, Y.; Yamamoto, R.; Tachibana, K. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform. In Proceedings of the ICASSP 202—023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Rhodes Island, Greece, 5 May 2023; pp. 1–5.
35. Morrison, M.; Kumar, R.; Kumar, K.; Seetharaman, P.; Courville, A.; Bengio, Y. Chunked Autoregressive GANforConditional Waveform Synthesis. In Proceedings of the International Conference on Learning Representations, Virtually, 25–29 April 2022.
36. Lorenzo-Trueba, J.; Drugman, T.; Latorre, J.; Merritt, T.; Pu trycz, B.; Barra-Chicote, R.; Moinet, A.; Aggarwal, V. Towards Achieving Robust Universal Neural Vocoding. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 181–185.
37. Zaïdi, J.; Seuté, H.; van Niekerc, B.; Carbonneau, M.-A. Daft-Exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis. *arXiv* **2021**, arXiv:2108.02271.
38. Ziyin, L.; Hartwig, T.; Ueda, M. Neural networks fail to learn periodic functions and how to fix it. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1583–1594.
39. Li, Y.A.; Han, C.; Jiang, X.; Mesgarani, N. HiFTNet: A fast high-quality neural vocoder with harmonic-plus-noise filter and inverse short time fourier transform. *arXiv* **2023**, arXiv:2309.09493.
40. Takida, Y.; Liao, W.H.; Uesaka, T.; Takahashi, S.; Mitsufuji, Y. Preventing posterior collapse induced by oversmoothing in gaussian vae. *arXiv* **2021**, arXiv:2102.08663.
41. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2020**, *33*, 6840–6851.
42. Ito, K.; Johnson, L. The LJ Speech Dataset. 2017. Available online: <https://keithito.com/LJ-Speech-Dataset/> (accessed on 14 May 2025).
43. Databaker: Databaker Technology. Chinese Standard Female Voice Database. Available online: <https://en.data-baker.com/datasets/freeDatasets/> (accessed on 14 May 2025).
44. ITU-T Recommendation P.800; Subjective Testing Methods for Voice Quality Assessment. ITU-T: Geneva, Switzerland, 1996.
45. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Victoria, BC, Canada, 19–21 May 1993; pp. 149–152.
46. Sundermann, D.; Jaitly, N. Real-time voice conversion with recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5450–5454.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.