

DiffStyleTTS: Diffusion-based Hierarchical Prosody Modeling for Text-to-Speech with Diverse and Controllable Styles

Jiaxuan Liu¹, Zhaoci Liu¹, Yajun Hu², Yingying Gao³, Shilei Zhang³, Zhenhua Ling^{*1},

¹NERCSLIP, University of Science and Technology of China, Hefei, China,

²iFLYTEK CO.LTD., Hefei, China,

³China Mobile Research Institute, Beijing, China

{jxliu, zcliu8}@mail.ustc.edu.cn, yjhu@iflytek.com,

{gaoyingying, zhangshilei}@chinamobile.com, zhling@ustc.edu.cn

Abstract

Human speech exhibits rich and flexible prosodic variations. To address the one-to-many mapping problem from text to prosody in a reasonable and flexible manner, we propose DiffStyleTTS, a multi-speaker acoustic model based on a conditional diffusion module and an improved classifier-free guidance, which hierarchically models speech prosodic features, and controls different prosodic styles to guide prosody prediction. Experiments show that our method outperforms all baselines in naturalness and achieves superior synthesis speed compared to three diffusion-based baselines. Additionally, by adjusting the guiding scale, DiffStyleTTS effectively controls the guidance intensity of the synthetic prosody.

1 Introduction

Speech synthesis, also known as text-to-speech (TTS), aims to turn text into almost human-like audio. Currently, most TTS models consist of three main components: a text analysis front-end, an acoustic model, and a vocoder. Among them, the naturalness and prosodic performance of speech primarily depend on the design of the acoustic model.

The acoustic model, at the heart of TTS, can be categorized as autoregressive and non-autoregressive. Autoregressive acoustic models, like Tacotron (Wang et al., 2017) and Transformer TTS (Li et al., 2019), have issues with word skipping, repeated reading and inference time increasing linearly with length of the Mel-spectrogram. Non-autoregressive acoustic models, like FastSpeech2 (Ren et al., 2021), excel in rapidly synthesizing high-quality speech. However, they are constrained by using a simple regression objective function for optimization, lacking probabilistic modeling, and the unimodal characteristics of

Gaussian distribution don't conform to the true distribution of acoustic features, which affects the prediction accuracy. The mean of the distribution also results in the problem of over-smoothing predictions, which restricts the diversity of generated prosodic features. These issues lead to weak fluctuations and unnaturalness in prosodic transfer and control tasks. Additionally, traditional prosodic transfer methods like Global Style Tokens (GST) (Wang et al., 2018) lack controllability over the intensity of prosodic transfer.

The recently emerged diffusion model has significant advantages in describing the complex distribution of high-dimensional and multi-modal features. In particular, the guidance of the conditional diffusion model (Dhariwal and Nichol, 2021) can well control the results. It effectively addresses issues like over-smoothing predictions and a lack of diversity through multi-step sampling. Currently, acoustic models based on the diffusion model, such as Diff-TTS (Jeong et al., 2021), Grad-TTS (Popov et al., 2021), DiffSinger (Liu et al., 2022), Guided-TTS (Kim et al., 2022), ProDiff (Huang et al., 2022), CoMoSpeech (Ye et al., 2023), etc., primarily use the Mel-spectrogram as the prediction target. There have been limited studies on predicting speech prosodic features via the conditional diffusion model. DiffProsody (Oh et al., 2024) is a diffusion-based prosody prediction model that constrains prosodic features through a discriminator, but it still lacks controllability over prosody during inference. In summary, the flexible transfer and control of speech prosody still remains underexplored.

Therefore, we propose a novel acoustic model, DiffStyleTTS, based on a conditional diffusion module and an improved classifier-free guidance (Ho and Salimans, 2021). It hierarchically models prosodic features using both coarse-grained style conditions and fine-grained prosodic descriptions, balances the diversity and quality of prosody

*Corresponding author. This work was funded by the National Nature Science Foundation of China under Grant U23B2053.

via classifier-free guidance, and is applied in prosodic transfer and control tasks to guide prosody prediction. Additionally, we introduce the dynamic thresholding method to address the issue of phoneme distortion caused by an excessive guiding scale. DiffStyleTTS is also designed to support various inference modes. Experiments¹ show DiffStyleTTS achieves higher naturalness and similar or faster synthesis speed compared to FastSpeech2 and other diffusion-based baselines. Compared to the FastSpeech2+GST and DiffProsody baselines, it demonstrates superior prosodic transfer capability, enabling flexible combination of speaker and prosodic features, alongside controllable prosodic transfer intensity via a guiding scale.

2 DiffStyleTTS

In this section, we propose DiffStyleTTS, a multi-speaker acoustic model that employs hierarchical prosody modeling and utilizes FastSpeech2 as its backbone. As shown in Figure 1, the encoder and decoder use the feed-forward Transformer (FFT) of FastSpeech2, along with a 5-layer convolutional PostNet (Shen et al., 2018) in the decoder. We use an embedding lookup table to capture the unique vocal characteristics of each speaker and a HiFi-GAN vocoder (Kong et al., 2020) to synthesize speech waveforms. The main modifications are replacing FastSpeech2’s original variance adaptor with a conditional diffusion module for hierarchical prosody modeling and introducing a GST module for style control.

2.1 Hierarchical Prosody Modeling

The DiffStyleTTS achieves hierarchical prosody modeling by considering prosodic features at two levels: coarse-grained implicit style conditions and fine-grained explicit prosodic descriptions. Implicit style conditions encompass broad descriptions of entire sentences, which are difficult to define intuitively and are encoded from the Mel-spectrogram during the training of the whole acoustic model. Explicit prosodic features include fine-grained prosodic descriptions of phonemes, such as pitch, energy, and duration, which can be directly and easily extracted from speech waveforms.

In DiffStyleTTS, the method of GST (Wang et al., 2018) is adopted to extract implicit style conditions from audio as shown in Figure 1. The implicit style conditions are decoupled into style

vectors corresponding to a group of global style tokens. Furthermore, explicit prosodic features are predicted using a conditional diffusion module, which introduces text embeddings and implicit style conditions as its conditional terms.

2.2 The Conditional Diffusion Module

The conditional diffusion module is guided with implicit style conditions to predict explicit prosodic features that align with it. The guided generation of the conditional diffusion module can be usually categorized into two approaches: classifier guidance (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2021). Classifier guidance requires an additional classifier, which slows down the inference speed, and its quality impacts the effectiveness of category generation. Therefore, DiffStyleTTS employs classifier-free guidance, which can avoid these issues as it doesn’t require direct calculation of the classifier gradient.

First, as shown in Figure 2, based on the mathematical principles of DDPM (Ho et al., 2020), the diffusion process of explicit prosodic features is defined by a fixed Markov chain from the initial data \mathbf{x}_0 to the latent variable \mathbf{x}_t as

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (1)$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (2)$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, $t = 0, 1, \dots, T$, and T is the step size. When adding a small Gaussian noise at each step, the module selects a small positive constant β_t from a variance table, which we define as a cosine schedule (Nichol and Dhariwal, 2021) to prevent rapid noise accumulation from linear addition. Then, the reverse process is also defined by a Markov chain from \mathbf{x}_t to \mathbf{x}_0 parameterized by the θ as

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad (3)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \quad (4)$$

which shows the step-by-step denoising of an isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to restore the original data \mathbf{x}_0 .

To guide the conditional diffusion module’s output using classifier-free guidance, two denoisers with identical architectures are designed to employ

¹<https://xuan3986.github.io/DiffStyleTTS/>

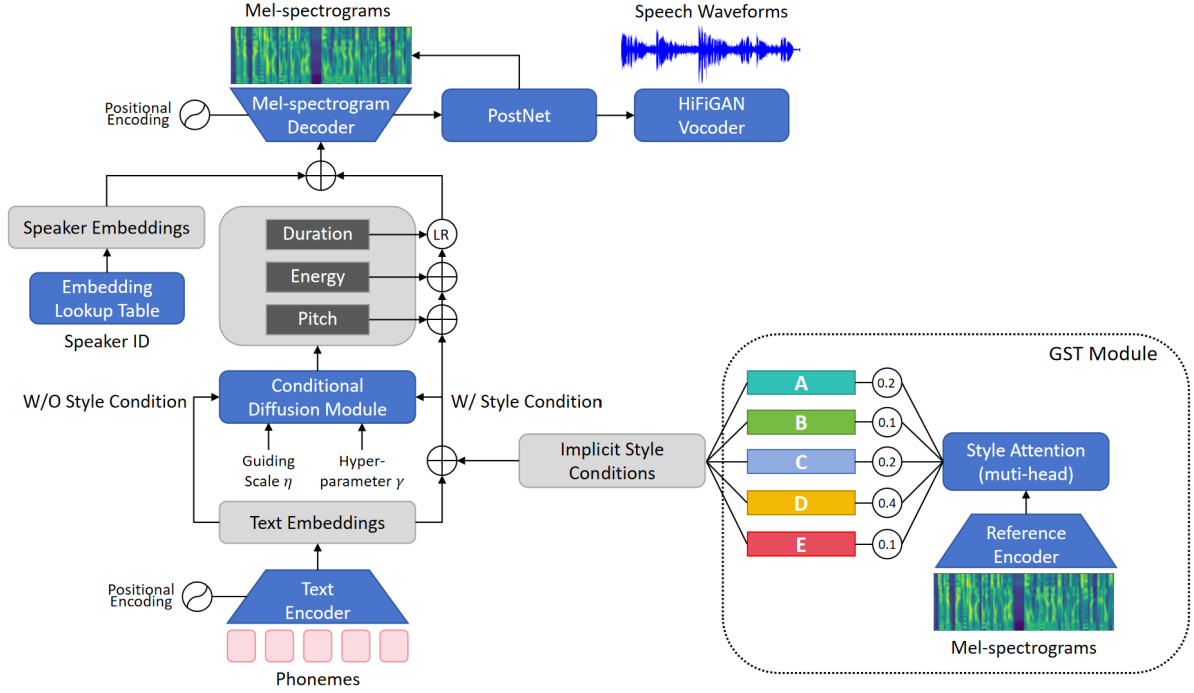


Figure 1: The model architecture of DiffStyleTTS. The LR refers to length regulator. The conditional diffusion module includes two denoisers $\Psi_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$ and $\Psi_{\theta_2}(\mathbf{x}_t, t, \mathbf{y})$.

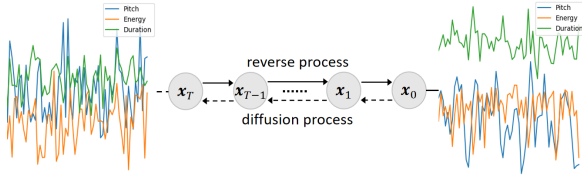


Figure 2: The diffusion process and reverse process of explicit prosodic features.

text embeddings \mathbf{y} as a condition to learn the mapping of phonemes into explicit prosodic features, with and without implicit style conditions \mathbf{c} respectively. The denoiser $\Psi_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$ uses \mathbf{y} and \mathbf{c} as the conditional input ($\text{condition} = \mathbf{y} + \mathbf{c}$), while the denoiser $\Psi_{\theta_2}(\mathbf{x}_t, t, \mathbf{y})$ leaves \mathbf{c} empty, using only \mathbf{y} as the conditional input. At each denoising step, the input of each denoiser consists of explicit prosodic features \mathbf{x}_t linearly combined with noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To model this noise, two denoisers are trained using the following training objective

$$\min_{\theta_1} L_{diff_c}(\theta_1) = \mathbb{E}_{\epsilon, \mathbf{x}_t, t, \mathbf{y}, \mathbf{c}} \|\epsilon - \epsilon_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})\|_2^2, \quad (5)$$

$$\min_{\theta_2} L_{diff_nc}(\theta_2) = \mathbb{E}_{\epsilon, \mathbf{x}_t, t, \mathbf{y}} \|\epsilon - \epsilon_{\theta_2}(\mathbf{x}_t, t, \mathbf{y})\|_2^2, \quad (6)$$

where we get noise outputs $\epsilon_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$ and $\epsilon_{\theta_2}(\mathbf{x}_t, t, \mathbf{y})$.

During inference, the two noise outputs are linearly interpolated to obtain the guided results

$$\tilde{\epsilon}_{\theta_1, \theta_2}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c}) = \epsilon_{\theta_2}(\mathbf{x}_t, t, \mathbf{y}, \emptyset) + \eta(\epsilon_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c}) - \epsilon_{\theta_2}(\mathbf{x}_t, t, \mathbf{y}, \emptyset)). \quad (7)$$

Here, η is the guiding scale used to adjust the guidance intensity, balancing the diversity and quality of explicit prosodic features.

Preliminary experiments found that when η was too high ($\eta \geq 7.0$), phoneme distortion occasionally occurred in some phonemes. These phonemes exhibited noise or elongation phenomena, resembling the ‘‘overexposed’’ issue from the previous study (Lin et al., 2024). To fix this, we improve classifier-free guidance via dynamic thresholding method, correcting the standard deviation of guidance results at each sampling step

$$\sigma_{cond} = std(\epsilon_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})), \quad (8)$$

$$\sigma_{cfg} = std(\tilde{\epsilon}_{\theta_1, \theta_2}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})), \quad (9)$$

$$\tilde{\epsilon}_{final} = \gamma \tilde{\epsilon}_{rescaled}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c}) + (1 - \gamma) \tilde{\epsilon}_{\theta_1, \theta_2}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c}). \quad (10)$$

This method corrects the standard deviation of $\tilde{\epsilon}_{\theta_1, \theta_2}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$ to the original standard deviation of $\epsilon_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$. A correction scale γ adjusts the

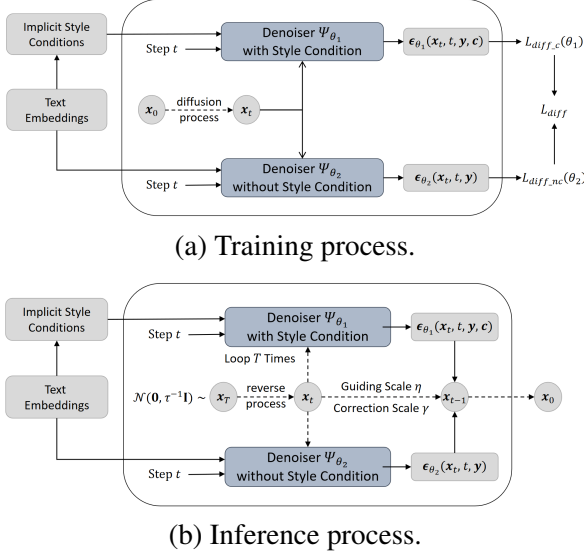


Figure 3: An illustration of the training and inference processes of the conditional diffusion module based on classifier-free guidance.

intensity of the correction to achieve the final corrected result $\tilde{\epsilon}_{final}$.

In summary, the conditional diffusion module can employ implicit style conditions c to guide the generation of explicit prosodic features, and use the guiding scale η and the correction scale γ to flexibly adjust the diversity of the explicit prosodic features and the guidance intensity.

2.3 Training and Inference

2.3.1 Training

Referring to Figure 1 and Figure 3(a), processed phonemes from the text analysis front-end are fed into the text encoder to generate text embeddings. This, along with implicit style conditions from the GST module, is then input into the conditional diffusion module, which includes two denoisers $\Psi_{\theta_1}(x_t, t, y, c)$ and $\Psi_{\theta_2}(x_t, t, y)$ trained via Eq. (5)(6). During training, the log scales of raw phoneme-wise pitch and duration, and raw phoneme-wise energy, are targeted for sampling. The guiding scale η and the correction scale γ are not involved in the training process. We add the implicit style conditions and the embeddings of raw pitch and energy to the text embeddings, then employ the length regulator to align the length based on raw duration. Frame-wise speaker embeddings are added before feeding into the decoder. Finally, decoded Mel-spectrograms are converted into speech waveforms using the pre-trained HiFi-GAN vocoder. The total loss function includes the

diffusion module loss, the loss of decoding Mel-spectrograms, and the residual loss of PostNet

$$L_{total} = L_{diff_c}(\theta_1) + L_{diff_nc}(\theta_2) + L_{decoder} + L_{mel}. \quad (11)$$

2.3.2 Inference

Referring to Figure 1 and Figure 3(b), three main inference modes are designed based on the trained DiffStyleTTS model.

(1) *Diversified controllable inference*. By tuning the guiding scale η and the correction scale γ , we can adjust the diversity and guidance intensity of explicit prosodic features, achieving diversified and controllable prosody prediction.

(2) *Prosodic transfer inference*. Given a reference utterance and a specified speaker ID, prosodic features are transferred from the reference utterance to this speaker. By tuning the guiding scale η and the correction scale γ , we can adjust the intensity of prosodic transfer.

(3) *Prosodic control inference*. Given a specified speaker ID and a token ID in the GST module, we can set the weights of other tokens to 0 and the weight of this token to 1, synthesizing prosody controlled only by that token. Besides, we allow for the flexible combination of style token weights, enabling the enhancement or diminishment of certain prosodic style. We can also scale the pitch, energy, and duration by multiplication with scaling factors to control prosodic values.

Additionally, we introduce a temperature hyperparameter τ (Popov et al., 2021) to sample terminal condition x_T from $\mathcal{N}(0, \tau^{-1}\mathbf{I})$ instead of x_T from $\mathcal{N}(0, \mathbf{I})$. Previous work has found that tuning τ can help to improve the quality of output.

3 Experiments

3.1 Experimental Setup

3.1.1 Dataset

We evaluated the proposed DiffStyleTTS model using a 54-hour private Mandarin Chinese dataset comprised of recordings from 9 male speakers of different ages, all of which belonged to the genres of novel, narration or story reading. Our dataset has high recording quality and diverse prosodic styles. We randomly sampled 20 utterances from each speaker’s recordings, and the total 180 utterances were reserved for validation and test, while the rest were used for training. All phoneme durations were extracted by an internal forced align-

ment tool based on HMM. Given phoneme boundaries, phoneme-wise pitch and energy features were obtained by averaging frame-wise pitch and energy. The frame-wise pitch was extracted using STRAIGHT (Kawahara, 1997) and interpolated at unvoiced frames, and the frame-wise energy was computed as the L2-norm of the amplitude spectrum derived using short-time Fourier transform. To get the input of the text encoder, each phoneme was represented as the concatenation of a phoneme identity embedding, a tone embedding and a positional embedding.

3.1.2 Model Configuration

The encoder encoded phonemes to 256-D text embeddings using 4 FFT blocks. while the decoder used 6 FFT blocks. The 5-layer convolutional PostNet in the decoder is comprised of 512 filters with shape 5×1 with batch normalization, followed by tanh activations on all but the final layer. The architecture of two denoisers utilized a bidirectional dilated convolution (Kong et al., 2021) similar to WaveNet (Oord et al., 2016), for predicting waveform signals. It consists of a stack of 12 residual layers, each layer with residual channels $C = 3$ and a kernel size of 3. The input tensor had a shape of $[B, C, L]$, where B was the batch size, C was the residual channels, and L was the length of phonemes. In the GST module, the token embeddings size was set to 256 and the token size was configured to 10. the multi-head attention with 4 attention heads used a softmax activation to output weights over the tokens.

3.2 Performance of Synthetic Speech

Subjective and objective evaluations were conducted to evaluate the performance of speech synthesized using DiffStyleTTS. In addition to the FastSpeech2 (Ren et al., 2021) baseline, a FastSpeech2 model with ground truth phoneme-wise prosodic features, a Grad-TTS (Popov et al., 2021) model, a Guided-TTS (Kim et al., 2022) model and a DiffProsody (Oh et al., 2024) model were also built for comparison. First, the naturalness mean opinion scores (MOS) of all models were evaluated by a listening test. A total of 14 participants evaluated 18 utterances for each model, selecting two samples from each speaker. Second, the accuracy of predicted prosodic probability distributions were evaluated using Jensen-Shannon (JS) Divergence. Third, the efficiency of different models were evaluated using the real time factor

Model	MOS \uparrow	JS Divergence \downarrow			RTF \downarrow
		Pitch	Energy	Duration	
Ground Truth	4.55 \pm 0.05	-	-	-	-
FastSpeech2	3.85 \pm 0.06	0.121	0.037	0.097	0.019
FastSpeech2*	4.11 \pm 0.07	-	-	-	-
Grad-TTS	4.08 \pm 0.07	0.115	0.040	0.088	0.250
Guided-TTS	4.15 \pm 0.07	0.080	0.033	0.050	0.479
DiffProsody	4.10 \pm 0.06	0.083	0.030	0.046	0.063
DiffStyleTTS	4.18\pm0.06	0.065	0.030	0.045	0.048
w/o ISC	3.92 \pm 0.07	0.090	0.038	0.078	-
w/o AISC	3.80 \pm 0.06	0.071	0.033	0.051	-
w/o TEC	2.05 \pm 0.24	0.445	0.125	0.390	-

Table 1: The naturalness MOS, JS Divergence and RTF of different models. The FastSpeech2* refers to FastSpeech2 with ground truth phoneme-wise prosodic features, the ISC refers to implicit style conditions, the TEC refers to text embedding conditions, and the AISC refers to adding implicit style conditions into text embeddings.

η	MOS \uparrow	CV \uparrow		
		Pitch	Energy	Duration
1.0	4.23 \pm 0.06	7.75	3.04	5.46
3.0	4.09 \pm 0.08	8.52	3.20	4.87
5.0	3.87 \pm 0.07	12.04	4.25	8.69
7.0	3.79 \pm 0.08	16.63	5.81	8.94

Table 2: The naturalness MOS, CV (%) of different guiding scale η in DiffStyleTTS.

(RTF). Fourth, different settings of η and γ used only for prosodic transfer and control can affect the MOS of DiffStyleTTS, as shown in Table 2. Therefore, we select the optimal configuration in this section. We employed the *diversified controllable inference* mode with $\eta = 1.0$ and $\gamma = 0.7$ in DiffStyleTTS. In all experiments, the step size T of all diffusion models was set to 200 to ensure rigor. Additionally, we ran multiple trials from different terminal conditions x_T to sample various explicit prosodic features, and then averaged scores for final evaluation results.

The results are shown in the first six rows of Table 1. We can see that our proposed DiffStyleTTS outperformed all baselines in naturalness and JS Divergence. DiffStyleTTS also achieved faster synthesis speed than the two diffusion models using Mel-spectrograms as modeling targets.

3.3 Prosodic Control

To study the effect of the guiding scale η in classifier-free guidance on balancing prosodic diversity and quality, we selected an utterance with pronounced prosodic variations and synthesized it using the *diversified controllable inference* mode

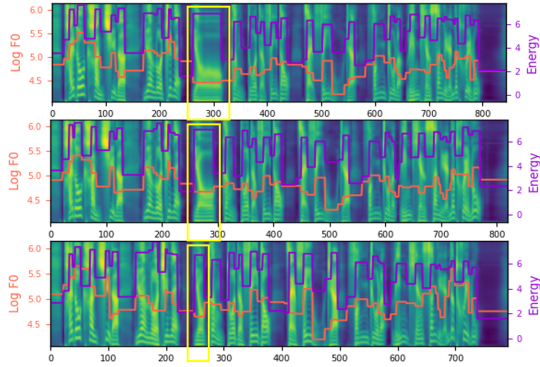


Figure 4: The Mel-spectrograms of a sentence synthesized with $\gamma = 0$ (top), 0.4 (middle) and 0.7 (bottom) respectively.

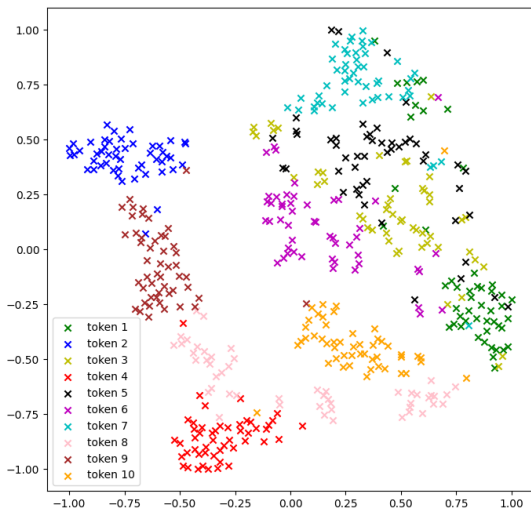


Figure 5: The t-SNE visualization results.

with $\gamma = 0.7$. We used naturalness MOS to evaluate the quality, and then calculated the coefficient of variation (CV) of phoneme-wise pitch, energy and duration to evaluate the diversity. A total of 12 participants evaluated four values of η to illustrate the effect of the guiding scale. The evaluation results, presented in Table 2, indicate that as the η increases, the diversity of explicit prosodic features increases while the audio quality decreases. Notably, when $\eta \geq 7.0$, phoneme distortion occasionally occurred.

To study the effect of the correction scale γ in the dynamic thresholding method, we randomly selected a test sentence, set the guiding scale $\eta = 7.0$ to cause phoneme distortion, and employed the *diversified controllable inference* mode to synthesize it three times with different values of γ . Figure 4 shows a distorted phoneme within the yellow box. When γ was set to 0 (top), the phoneme distortion was pronounced, exhibiting noticeable elongation.

This issue was alleviated at $\gamma = 0.4$ (middle) and effectively resolved at $\gamma = 0.7$ (bottom).

To verify the ability of different tokens in the GST module to control the generation of explicit prosodic features, we randomly selected 50 sentences from the evaluation set to predict explicit prosodic features. During inference, we employed the *prosodic control inference* mode to synthesize a total of 500 explicit prosodic samples, which included phoneme-wise pitch, energy, and duration. We then computed the mean, standard deviation, median, minimum, maximum, skewness, and kurtosis for each sample, resulting in a 21-dimensional prosodic distribution vector for each sample. We used t-SNE (van der Maaten and Hinton, 2008) to reduce the dimensions of these vectors to two dimensions for visualization.

Figure 5 shows that these 500 samples can be well divided into 10 clusters, each matching one of the 10 tokens. which indicates the effectiveness of the hierarchical prosody modeling in DiffStyleTTS on controlling the prosodic features of synthetic speech via implicit style conditions.

3.4 Prosodic Transfer

To evaluate the effect of DiffStyleTTS in prosodic transfer, we used FastSpeech2+GST, i.e., incorporating a GST module before the variance adaptor in FastSpeech2, and the DiffProsody as the baselines, we selected one reference utterance each from two speakers (A and B) with distinct prosodic styles. Then, we randomly selected two sentences each from the other eight speakers, excluding the reference speakers, resulting in a total of 16 sentences for prosodic transfer. The *prosodic transfer inference* mode was used to transfer the prosody of the reference utterance to the other eight speakers. We conducted a subjective preference (%) test involving 12 participants to compare the two models and calculated the p-value of t-test to assess significance of differences. As shown in Table 3, DiffStyleTTS significantly outperformed two baselines on this prosodic transfer task ($p < 0.05$). We also investigated the effect of the guiding scale η on the intensity of prosodic transfer, setting $\gamma = 0.7$. The experimental results in Figure 6 compare the transfer effects of three different guiding scales. It’s observed that as η increased, the prosodic transfer effect became more pronounced. We recommend listening to the examples on the demo page ¹.

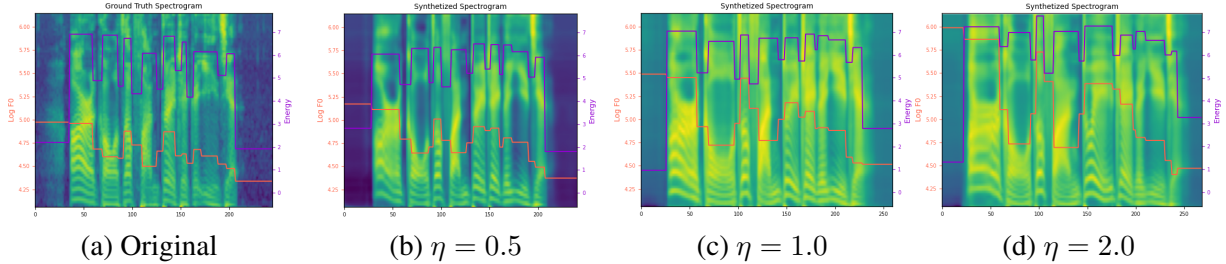


Figure 6: By tuning the guiding scale η , we adjusted the intensity of the prosodic transfer from reference audio to original audio (a) to synthesize three Mel-spectrograms (b) (c) and (d).

	Preference			p-value
	FastSpeech2+GST	Neutral	DiffStyleTTS	
Reference A	33.13	14.37	52.50	0.0077
Reference B	31.25	21.87	46.88	0.0249

	Preference			p-value
	DiffProsody	Neutral	DiffStyleTTS	
Reference A	16.67	33.33	50.00	0.0134
Reference B	18.13	29.37	52.50	0.0231

Table 3: Subjective preference results for prosodic transfer.

3.5 Ablation Studies

We conducted ablation studies to demonstrate the effectiveness of key components in DiffStyleTTS. The results are presented in the last three rows of Table 1.

To verify the guiding effect of implicit style conditions on explicit prosodic features, we can observe that removing the implicit style conditions \mathbf{c} from the denoiser $\Psi_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$, prohibiting the use of classifier-free guidance, i.e., w/o ISC, but retaining them added to text embeddings, led to a decrease in MOS and an increase in JS divergence. When the implicit style conditions added to the text embeddings were removed and only retained as the condition \mathbf{c} in $\Psi_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$, i.e., w/o AISC, we can also observe a decrease in MOS, which verifies the implicit style conditions contain prosodic features. Furthermore, removing the text embedding conditions \mathbf{y} from the denoiser $\Psi_{\theta_1}(\mathbf{x}_t, t, \mathbf{y}, \mathbf{c})$, i.e., w/o TEC, resulted in poor quality of synthetic prosody, indicating the importance of text embedding conditions on prosodic alignment. In summary, experiments show that these key components contributed significantly to the performance of DiffStyleTTS.

4 Conclusion

This paper proposes a multi-speaker acoustic model, DiffStyleTTS, based on a conditional diffusion module and an improved classifier-free guid-

ance. We hierarchically model prosodic features at both implicit and explicit levels. Text embeddings and implicit style conditions are combined as the diffusion module’s conditions. To predict explicit prosodic features, the dynamic thresholding method is employed to improve classifier-free guidance and then adjust the guidance intensity. Experiments show that our proposed model achieves higher naturalness compared to all baselines, and faster synthesis speed compared to diffusion-based baselines. Additionally, DiffStyleTTS demonstrates superior prosodic transfer capabilities and flexibility in comprehensive prosodic transfer and control.

5 Limitation

Although our work on prosody prediction has made it more flexible and controllable, we haven’t successfully decoupled prosody from speaker timbre. In addition, we can also observe some overlapping samples after dimensionality reduction, indicating that the tokens are not entirely independent and share common prosodic styles. This suggests that the implicit style conditions need further classification. To achieve better disentanglement of prosody will be a task for our future work.

References

- Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Proc. NeurIPS*, pages 8780–8794.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proc. NeurIPS*.
- Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *Proc. NeurIPS*.
- Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. Prodiff: Progressive fast diffusion model for high-quality text-to-speech.

- In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605. ACM.
- Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. 2021. Diff-tts: A denoising diffusion model for text-to-speech. In *Proc. Interspeech 2021*, pages 3605–3609. ISCA.
- Hideki Kawahara. 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proc. ICASSP*, volume 2, pages 1303–1306. IEEE.
- Heeseung Kim, Sungwon Kim, and Sungroh Yoon. 2022. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *Proc. ICML*, volume 162, pages 11119–11133. PMLR.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*.
- Zhifeng Kong, Wei Ping, Jiayi Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *Proc. ICLR*. OpenReview.net.
- Naihan Li, Shujie Liu 0001, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proc. AAAI*, volume 33, pages 6706–6713. AAAI Press.
- Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024. Common diffusion noise schedules and sample steps are flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5392–5399. IEEE.
- Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proc. AAAI*, pages 11020–11028. AAAI Press.
- Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proc. ICML*, volume 139, pages 8162–8171. PMLR.
- Hyung-Seok Oh, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–13.
- Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop*, page 125. ISCA.
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proc. ICML*, volume 139, pages 8599–8608. PMLR.
- Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *Proc. ICLR*.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. ICASSP*, pages 4779–4783. IEEE.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.
- Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010. ISCA.
- Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proc. ICML*, volume 80, pages 5167–5176. JMLR.
- Zhen Ye, Wei Xue, Xu Tan, Jie Chen, Qifeng Liu, and Yike Guo. 2023. Comospeech: One-step speech and singing voice synthesis via consistency model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1831–1839. ACM.