Diverse and Expressive Speech Prosody Prediction with Denoising Diffusion Probabilistic Model

Xiang Li[1,†], Songxiang Liu[2], Max W. Y. Lam[2], Zhiyong Wu[1,3,], Chao Weng[2], Helen Meng[1,3]*

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China

`xiang-li20@mails.tsinghua.edu.cn`, {`shaunxliu`, `maxwylam`}`@tencent.com`,
`zywu@sz.tsinghua.edu.cn`, `cweng@tencent.com`, `hmmeng@se.cuhk.edu.hk`

Abstract

Expressive human speech generally abounds with rich and flexible speech prosody variations. The speech prosody predictors in existing expressive speech synthesis methods mostly produce deterministic predictions, which are learned by directly minimizing the norm of prosody prediction error. Its unimodal nature leads to a mismatch with ground truth distribution and harms the model's ability in making diverse predictions. Thus, we propose a novel prosody predictor based on the denoising diffusion probabilistic model to take advantage of its high-quality generative modeling and training stability. Experiment results confirm that the proposed prosody predictor outperforms the deterministic baseline on both the expressiveness and diversity of prediction results with even fewer network parameters.

Index Terms: expressive speech synthesis, prosody prediction, denoising diffusion probabilistic model

1. Introduction

Speech synthesis aims to convert a given text into the corresponding speech audio. In modern speech synthesis models, the naturalness and sound quality of the synthesized audios have been promoted to a human-like level by incorporating neural-network-based acoustic models and vocoders, which are now widely accessible in many real-world applications [1, 2, 3, 4]. On the other hand, how to reach the expressiveness and diversity of human speech in synthesized audio remains an open question. Thus, the task of expressive speech synthesis is drawing growing attention recently.

Existing expressive speech synthesis methods generally model the prosody in human speech as an additional condition to the text-to-speech (TTS) backbone to stylize the synthesized audio [5, 6, 7]. During training, the speech prosody representation is extracted from the target speech as the desired condition of the TTS backbone and the training target of a speech prosody predictor, which makes its prediction according to the input text [8]. During inference, given the input text, the prediction result of the trained speech prosody predictor is applied as the actual condition of the TTS backbone. As a result, the speech prosody prediction module is not only necessary for the inference process in real-world applications, but also the crucial component that directly affects the expressiveness and diversity of the prosody in synthesized speech.

Typical speech prosody predictors like the phoneme duration, fundamental frequency (pitch), and energy predictors in

FastSpeech2 [3] and FastPitch [9] take pre-extracted duration, pitch, and energy features as the target speech prosody representation, and utilize the output of the text encoder in TTS backbone as prediction conditions. Other methods manage to improve the prediction conditions by incorporating context information or semantic structure of the input text [10, 11]. The prosody predictors in these methods generally establish a deterministic mapping from input text to predicted prosody, and is trained via minimizing the L1 or L2 norm of prediction error, by assuming that the target prosody distribution is an independent unimodal Laplace or Gaussian distribution [12].

However, the prosody of human speech is extremely flexible. Changes in intonation, speaking rate, emphasis, context, and other factors provide numerous ways to utter the same text, resulting in multimodal and correlated prosody distributions. Consequently, the diversity and expressiveness of the prediction results from previous prosody predictors tend to be diminished by its deterministic characteristics and the over-smoothing effect due to simplified assumptions on prosody distribution.

Thus, we propose to replace the deterministic prosody predictors with an alternative based on the denoising diffusion probabilistic model (DDPM) [13]. DDPM and its variants have shown powerful capacity in generative modeling with respect to high-quality sampling, gorgeous model coverage, and sample diversity [14, 15, 16, 17, 18]. Moreover, DDPMs can be trained with a simplified variant of the evidence lower bound of data likelihood, leading to stable optimization. Based on this thinking, this paper focuses on investigating the use of DDPMs for modeling the diverse and flexible one-to-many mapping from text to prosody in human speech. Experiment results reveal that the proposed DDPM-based prosody predictor significantly outperforms the deterministic baseline on both the diversity and expressiveness of prediction results, indicating the effectiveness of introducing DDPM to the task of speech prosody prediction.

2. Methodology

As shown in Figure 1, we incorporate the proposed DDPM-based prosody predictor with a pre-trained TTS backbone based on FastSpeech2 to construct an expressive TTS system. Following FastSpeech2, we encode speech prosody into a 3-dimensional feature of pitch, energy, and duration.
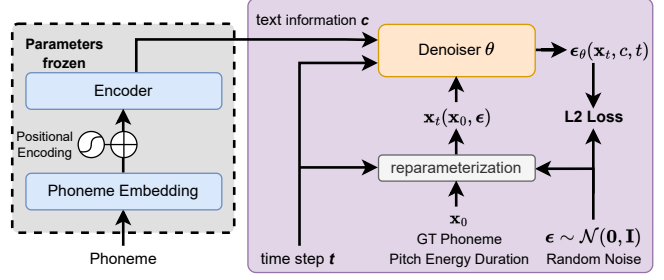
2.1. Text-to-speech backbone

The backbone of the proposed expressive TTS system is an adapted FastSpeech2 with a modified variance adaptor. The phoneme embedding, encoder, and mel spectrogram decoder

---

Figure 1: *The overall architecture of the expressive TTS system with the proposed DDPM-based prosody predictor*
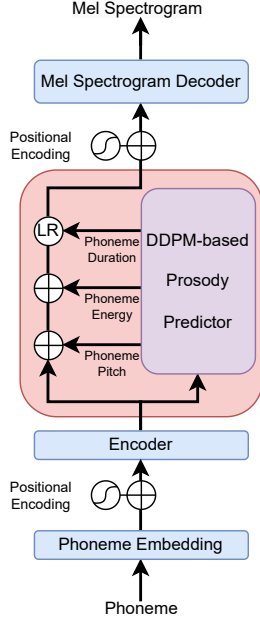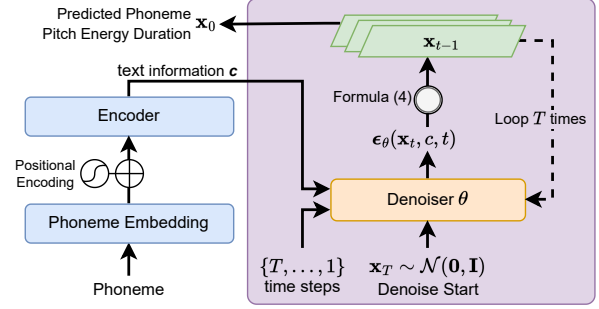
(a) *The FastSpeech2-based TTS backbone*

(b) *The training procedure of the DDPM-based prosody predictor*

(c) *The inference procedure of the DDPM-based prosody predictor*

modules are the same as those in FastSpeech2. While for the variance adaptor, instead of applying frame-wise pitch and energy features, we use phoneme-wise pitch and energy to improve training stability. The phoneme-wise pitch, energy, and duration values are provided by the proposed DDPM-based prosody predictor, as displayed in Figure 1a. Similar to the original setting, the given pitch and energy are first quantized and mapped to embedding vectors via a codebook, then added to the phoneme-wise output of the encoder as a stylized input condition, which is repeated with a length regulator (LR) module according to the given phoneme durations and fed to the mel spectrogram decoder to generate the final mel spectrogram.

2.2. DDPM-based prosody predictor

2.2.1. Preliminary knowledge on DDPM

DDPM is a generative model built upon a forward process that is fixed to a Markov chain that diffuses data $\mathbf{x}_0$ into white noise, and a reverse process that samples $\mathbf{x}_0$ from white noise via another Markov chain with learned Gaussian transitions.

Practically, the former process diffuses the $\mathbf{x}_{t-1}$ with a small Gaussian noise to obtain $\mathbf{x}_t$ at each diffusion step $t \in [1, T]$ according to the fixed variance schedule $\beta_1, \ldots, \beta_T$:

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

After T steps, $\mathbf{x}_0$ is transformed into $\mathbf{x}_T$.

The reverse process starts at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$, and is approximated via a neural denoiser θ as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

where $\sigma^2 = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, with $\alpha_t := 1-\beta_t$ and $\overline{\alpha} := \prod_{s=1}^t \alpha_s$. The parameters θ of the denoiser are trained via minimizing a variational bound of the negative log-likelihood:

$$\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right]$$

which is eventually simplified as the following variant by the reparameterizations in Formula 1 and 2 where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

$$L_{simple}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2\right]$$

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) \tag{2}$$

The complete proof of these formulas can be found in [13].

2.2.2. DDPM-based prosody predictor

The proposed prosody predictor is a DDPM modeled on 3-dimensional prosody features $\mathbf{x}_0$, which consists of phoneme-wise pitch, energy, and duration, respectively. We employ non-causal WaveNet as the underlying denoiser network θ [15, 19].

As shown in Figure 1b, during each training step of the prosody predictor, we first take the text information generated by the FastSpeech2 encoder as prediction condition c. Note that the parameters of the original FastSpeech2 modules are frozen the whole time to keep the TTS backbone intact. Then we uniformly sample the Markov chain time step t from $[1, T]$, based on which the pre-extracted ground truth (GT) phoneme prosody features $\mathbf{x}_0$ are diffused into $\mathbf{x}_t$ via the reparameterization in Formula 1. Given condition c, t and input $\mathbf{x}_t$, the denoiser θ outputs $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)$, and updates its parameters by the gradient in Formula 3.

$$\nabla_\theta\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, c, t)\|^2 \tag{3}$$

The inference procedure of the prosody predictor is depicted in Figure 1c. Given text information c, we first sample the denoise start $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then iteratively compute $\mathbf{x}_{t-1}$ based on $\mathbf{x}_t$ according to Formula 4 for $t = T, \ldots, 1$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ except for $\mathbf{z} = \mathbf{0}$ when $t = 1$.

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)\right) + \sigma_t\mathbf{z} \tag{4}$$

Table 1: *Overall performance comparison, including: (i) Mean Opinion Score (MOS) of speech expressiveness with 95% confidence intervals; (ii) Jensen Shannon (JS) Divergence between the prosody distributions of the prediction and ground truth; (iii) Number of network parameters in the prosody predictors.*

Systems	MOS (↑)	JS Divergence (↓)			#Params in prosody predictor
		Pitch	Energy	Duration	
FastSpeech2 with original prosody predictor	3.58 ± 0.07	0.199	0.056	0.119	1,185,027
FastSpeech2 with proposed prosody predictor	$\mathbf{3.98 \pm 0.08}$	**0.085**	**0.055**	**0.056**	$\mathbf{738,499}$
FastSpeech2 with ground truth prosody	4.23 ± 0.07	-	-	-	-
Ground truth audio (Reconstructed)	4.39 ± 0.06	-	-	-	-

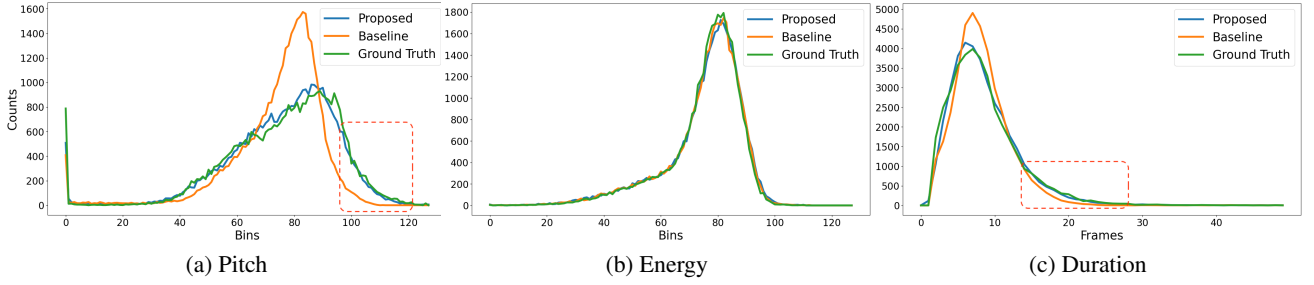(a) Pitch   (b) Energy   (c) Duration

Figure 2: *Distribution of the quantized prosody in predictions and ground truth for phoneme "ng". Red rectangles highlight that the baseline tends to avoid making aggressive predictions, resulting in over-smoothed distributions. (Orange lines correspond to the baseline, blue lines correspond to the proposed method, green lines correspond to the ground truth distribution)*
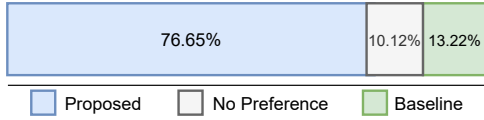

Figure 3: *AB preference test results*

After looping for T times, the final prosody prediction $\mathbf{x}_0$ is obtained. The $\mathbf{x}_0$ is subsequently sent to the modified variance adaptor in the TTS backbone and eventually guides the mel spectrogram decoder to synthesize speech stylized by the predicted prosody.

3. Experiments

3.1. Dataset and model details

We apply the proposed method to a private Mandarin audiobook dataset with 28.23 hours of speech audio voiced by a professional male speaker, since there is few public Mandarin corpus with diverse and expressive prosody variations, as well as satisfying sound quality. We randomly sample 5% of the dataset and reserve it for validation and testing, while the rest 95% is used for training. All of the phoneme duration is extracted by a pre-trained speech recognition model automatically. Based on which, the phoneme-wise pitch and energy features are obtained through averaging over frame-wise pitch and energy, where the frame-wise pitch is extracted via Crepe [20], and the frame-wise energy is computed as the L2-norm of the amplitude of each short-time Fourier transform frame. To boost the network's training, the prosody predictor's practical duration target is computed by taking the log scale of the raw phoneme duration.

The encoder and mel spectrogram decoder of the TTS backbone are made up of 4 and 6 Feed-Forward Transformer blocks

[21] with channel width 256, respectively. The variance adaptor quantizes input pitch and energy values into 128 bins. The proposed prosody predictor is built upon a DDPM with $T = 500$ and linear noise schedule $\beta_{1:T} = \{10^{-4}, \ldots, 0.06\}$, and the underlying non-causal WaveNet θ consists of 10 stacked residual layers with channel width 64. We also train the original prosody predictors in FastSpeech2 with mean squared error (MSE) as the deterministic **baseline** model for comparison, which contains 60.5% more network parameters, as shown in Table 1. The baseline predictors are trained along with the FastSpeech2 backbone with batch size 16 on an Nvidia GeForce RTX 2080Ti for 900,000 steps. The proposed model is trained under the same setting for the same number of steps, which took approximately 31 hours. The training and evaluation of both methods are performed in a single run, without grid searching for hyperparameters and random seeds. A pre-trained HiFi-GAN [4] is utilized as the backend for waveform reconstruction.

3.2. Expressiveness

To evaluate the expressiveness of the speech synthesized by the proposed method, we conduct a Mean Opinion Score (MOS) test and an AB preference test on 22 utterances from the test set, with 22 native speakers serving as subjects. In the MOS test, each subject is asked to rate the given audios on a scale of 1 to 5 regarding their expressiveness. As shown in Table 1, the proposed method reaches higher scores than the baseline. In the AB test, each subject is asked to choose the more expressive audio of 2 anonymous ones synthesized by the proposed method and baseline. As shown in Figure 3, the proposed method is favored with an overwhelming 76.65% of the preference, while the baseline only receives 13.22%.

The prevailing subjective test results on expressiveness are consistent with the statistics of the prosody prediction distribu-
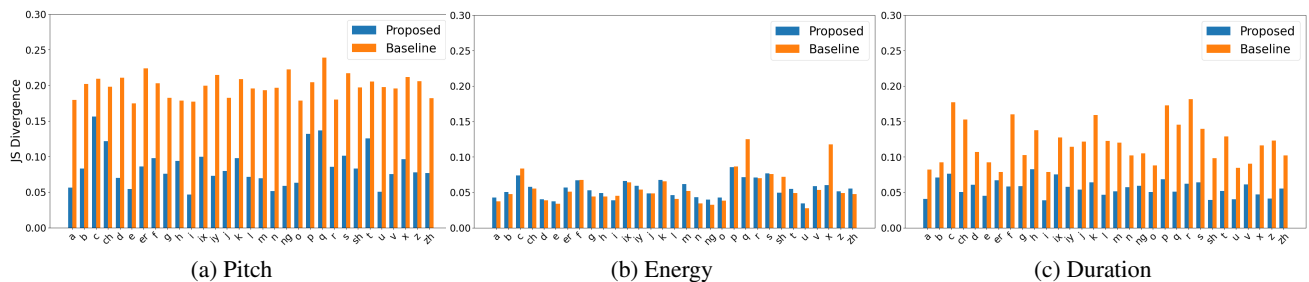
(a) Pitch   (b) Energy   (c) Duration

Figure 4: *Detailed values of JS Divergence between the predicted and ground truth prosody over each phoneme. Higher bars stand for larger divergence, indicating worse diversity of the prediction. (Orange bars correspond to the baseline, blue bars correspond to the proposed method)*
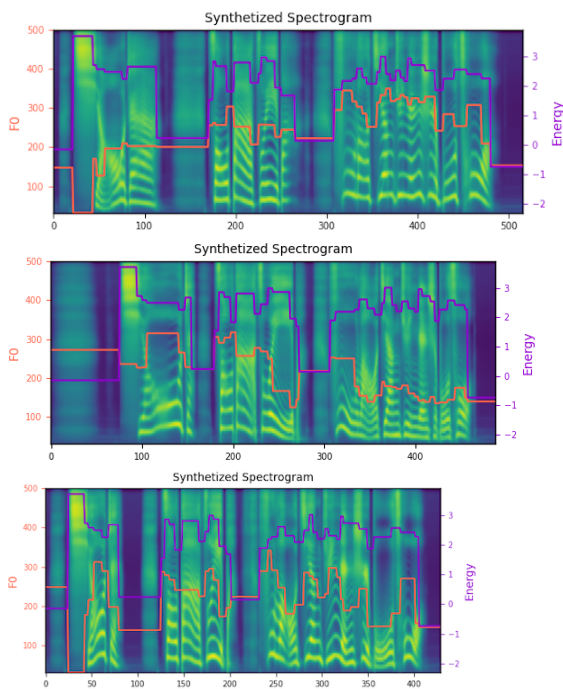

Figure 5: *Prosody prediction results generated by the proposed method in different runs on the same input text, stylizing the synthesized audio with diverse prosody.*

tion. As plotted in Figure 2, using the phoneme *"ng"* as an example, the predictions from the proposed method accumulate to a distribution closer to ground truth distribution. In contrast, the baseline tends to make over-smoothed predictions, resulting in squeezed distributions. From our empirical study, the ground truth distributions of the pitch and duration are more complex than that of the energy, which appears to have low variance in a narrow bell curve. This explains the defective performance of the baseline, as it is limited in unimodal predictions.

3.3. Diversity

To reveal the diversity of the predicted prosody, we compute the Jensen Shannon (JS) divergence between the distributions of the quantized prosody in predictions and ground truth to measure how well the predictions fit into the ground truth distribution [22]. Results in Table 1 show that the proposed prosody predictor achieves lower JS divergence on all of the 3 prosody features, indicating a better diversity of prediction results. We

also plot out the JS divergence values over each phoneme in Figure 4 to observe the characteristics of prediction diversity, which turns out to match the observation in Figure 2. For pitch and duration prediction, the proposed method produces better results on all the phonemes. For energy prediction, there is no significant difference on most phonemes except a few voiceless sounds like *"q"*, *"x"* and *"sh"*.

We can also directly demonstrate the diversity of the proposed method by sampling multiple times on the same text. As shown in Figure 5, the predicted prosody varies in different samples, resulting in diverse synthesized spectrograms. Meanwhile, the predicted prosody is still stably in coherence with the text, as reflected in Table 1 where no significant expansion of confidence interval in the MOS result is observed.

4. Conclusion and Discussion

In this work, we propose a DDPM-based speech prosody predictor and attach it to a FastSpeech2-based TTS backbone to form an expressive TTS system. Compared with traditional deterministic predictors, the proposed DDPM-based method enhances the expressiveness of synthesized speech and brings noticeable improvements to the diversity of the predicted prosody, which is still in coherence with the given text. Please refer to our demo website[1] for listening to samples.

Nevertheless, the iterative sampling process of the proposed method brings increased latency to the TTS system. We conduct a computation cost test on an Nvidia GeForce RTX 2080Ti, and find that the average real-time factor (RTF) of the proposed DDPM-based prosody predictor with full 500 denoising steps is 0.47, while the RTF of the original FastSpeech2 variance predictor is 0.31×10^{-3}. Although the DDPM-based approach achieves real-time inference, it is slowed down by the iterative sampling process. In the future, we will consider introducing fast sampling techniques that are able to produce results within a couple of denoising steps, as well as modifying the underlying DDPM with novel improvements that emerge in the latest diffusion models [23, 24].

5. Acknowledgements

---

[1]https://thuhcsi.github.io/interspeech2023-DiffVar

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 2017, pp. 4006–4010.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

[4] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020, p. 17022–17033.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[6] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[7] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards Multi-Scale Style Control for Expressive Speech Synthesis," in *Proc. Interspeech 2021*. Brno, Czech: ISCA, 2021, pp. 4673–4677.

[8] Z. Hodari, A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman, "Camp: a two-stage approach to modelling prosody in context," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6578–6582.

[9] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[10] Y. Ren, M. Lei, Z. Huang, S. Zhang, Q. Chen, Z. Yan, and Z. Zhao, "Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7577–7581.

[11] S. Lei, Y. Zhou, L. Chen, Z. Wu, S. Kang, and H. Meng, "Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7922–7926.

[12] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting over-smoothness in text to speech," in *The 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Dublin, Ireland: Association for Computational Linguistics, 2022, p. 8197–8213.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[14] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[15] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2021.

[16] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A Denoising Diffusion Model for Text-to-Speech," in *Proc. Interspeech 2021*, 2021, pp. 3605–3609.

[17] S. Liu, D. Su, and D. Yu, "Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans," *CoRR*, vol. abs/2201.11972, 2022.

[18] T. Tran, T. Nguyen, H. Bui, K. Nguyen, N. G. Vo, T. V. Pham, and T. Quan, "Naturalness improvement of vietnamese text-to-speech system using diffusion probabilistic modelling and unsupervised data enrichment," in *Intelligence of Things: Technologies and Applications*. Cham: Springer International Publishing, 2022, pp. 376–387.

[19] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[20] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[21] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[22] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[23] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations (ICLR)*, 2022.

[24] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *International Conference on Learning Representations (ICLR)*, 2022.