NoreSpeech: Knowledge Distillation based Conditional Diffusion Model for Noise-robust Expressive TTS

Dongchao Yang[1], Songxiang Liu[2], Helin Wang[3], Jianwei Yu[2], Chao Weng[2], Yuexian Zou[1]*

[1]Peking University, China
[2]Tencent AI Lab, China
[3]Johns Hopkins University, USA

dongchao98@stu.pku.edu.cn

Abstract

Expressive text-to-speech (TTS) can synthesize a new speaking style by imitating prosody and timbre from a reference audio, which faces the following challenges: (1) The highly dynamic prosody information in the reference audio is difficult to extract, especially, when the reference audio contains background noise. (2) The TTS systems should have good generalization for unseen speaking styles. In this paper, we present a **no**ise-**r**obust **e**xpressive TTS model (NoreSpeech), which can robustly transfer speaking style in a noisy reference utterance to synthesized speech. Specifically, our NoreSpeech includes several components: (1) a novel DiffStyle module, which leverages powerful probabilistic denoising diffusion models to learn noise-agnostic speaking style features from a teacher model by knowledge distillation; (2) a Vector Quantization (VQ) block, which maps the style features into a controllable quantized latent space for improving the generalization of style transfer; and (3) a straightforward but effective parameter-free text-style alignment module, which enables NoreSpeech to transfer style to a textual input from a length-mismatched reference utterance. Experiments demonstrate that NoreSpeech is more effective than previous expressive TTS models in noise environments.

Index Terms: Text-to-speech, style transfer, diffusion model, knowledge distillation, vector quantization

1. Introduction

Text-to-speech (TTS) [1, 2] aims to transform text into almost human-like speech, which attracts broad interest in the deep learning community. TTS is a key technology for spoken language understanding [3, 4, 5, 6]. Nowadays, TTS models have been extended to more complex scenarios, including multiple speakers timbre, emotions, and speaking styles for expressive synthesis [7]. Style modeling and transferring have been studied for decades in the TTS community: Wang *et al.* [8] proposed to use global style tokens to control and transfer the global style. Li *et al.* [9] adopt a multi-scale style encoder to assist synthesis expressive speech. Yang *et al.* [10] propose to use text prompt control speaking style. Min *et al.* [11] proposed Meta-StyleSpeech, which uses a meta-learning training strategy for multi-speaker TTS synthesis. Huang *et al.* [7] proposed a multi-level style adaptor to transfer speaking style. However, these methods assume that the reference audio is recorded in ideal environments (without noise interference). This assumption prevents expressive TTS models from being applied in many real-world scenarios, *e.g.* the reference audio recorded
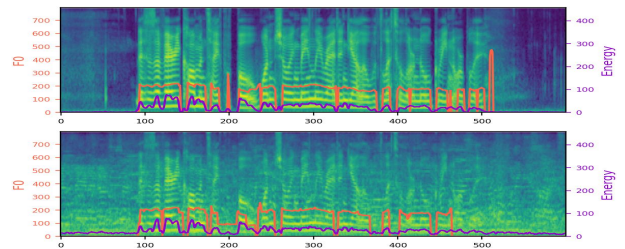
Figure 1: *The F0 and energy comparisons of clean speech (the first row) and noisy speech (the second row).*

by users may include noise. Zhang *et al.* [12] demonstrated that the fundamental frequency (F0) and energy can be affected by adding noise, which are key components of speaking style, as Figure 1 shows. To eliminate the effect of noise in reference audio, many methods have been proposed [13, 14, 15, 16]. These methods can be classified into two types: (1) using a pre-trained speech enhancement model to eliminate noise in reference audio [13], which heavily relies on the performance of a speech enhancement (SE) model; (2) decomposing the noise information via adversarial training [15] or information bottleneck [14, 16]. However, the speech enhancement can introduce unexpected distortions and the adversarial training and information bottleneck strategies need complex parameter setting and training tricks, which makes them hard to be widely applied.

In this paper, instead of extracting style information from noisy reference audio, we propose to reconstruct style information by learning relevant parameters for distribution modeling. Specifically, we propose a knowledge distillation based conditional diffusion model [17] that can directly generate style representation in a latent space conditioned on the noisy reference audio, named DiffStyle. Our DiffStyle is inspired by a popular diffusion-based speech enhancement model, CDiffuSE [18], which has shown that diffusion models are capable of restoring clean speech component from a noise-contaminated speech signal. The CDiffuSE model reconstructs clean waveform conditioned on a noisy spectrogram. The DiffStyle module in NoreSpeech is different from the CDiffuSE model in the following aspects: (1) Instead of generating highly stochastic time-domain audio signal (1 second composed of thousands of sample points), DiffStyle works on the more compressed frame-level features. (2) The CDiffuSE model reconstructs complete speech signal in time domain, while DiffStyle learns to generate prosody-related style features supervised by a pre-trained teacher model. Furthermore, we explore two aspects of improving the generalization of style transfer: (1) To transfer style to a textual input from a length-mismatched reference utterance,
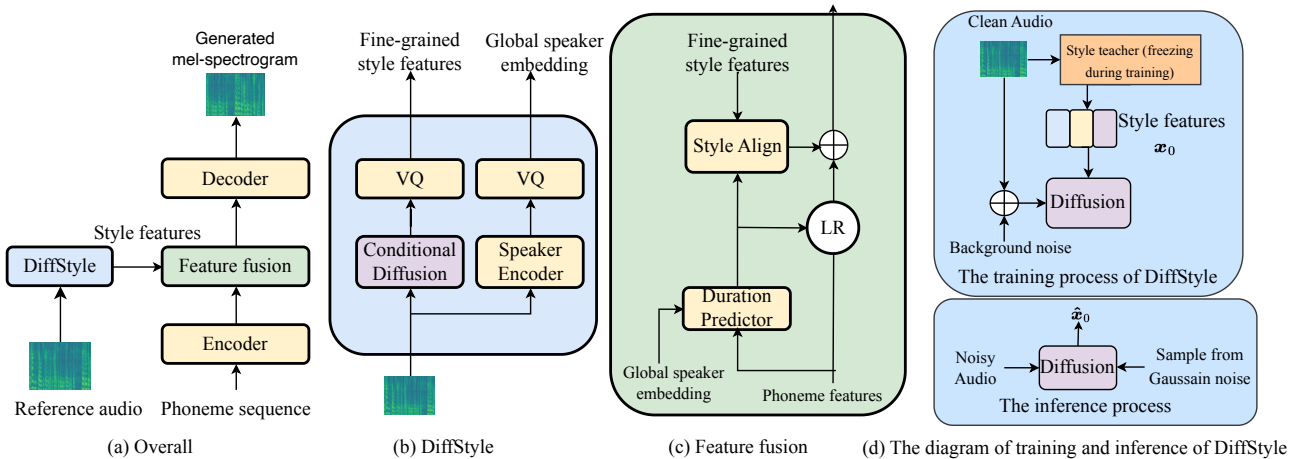
Figure 2: *(a) shows the overview of our proposed NoreSpeech. (b) and (d) show the details of DiffStyle. In subfigure (c), LR denotes the length regulator, the decoder includes 4 Transformer blocks and one flow-based post-net, which is the same as [7]. Subfigure (d) shows the details of training and inference stages of DiffStyle. Note that the style teacher model is freezing during the training process.*

we propose a parameter-free style-alignment module; (2) To transfer unseen speaking styles, we use a Vector Quantization (VQ) module [19] to map the style features into a controllable latent space, which has previously been shown to be an effective method [7]. Experimental results show that our proposed NoreSpeech has better performance than baselines in noise environments. In the following, we first introduce the details of NoreSpeech, then we present the experiments in Section III. We give the conclusion of this study in Section IV.

2. Proposed method

In this section, we first formulate the noise-robust expressive TTS model for style transfer. Then we overview our proposed **noi**se-**r**obust **e**xpressive TTS model (NoreSpeech), following which we introduce several critical components, including the style teacher model, DiffStyle, and feature fusion module.

2.1. Problem formulation

Style transfer aims to generate high-quality and similarity speech samples with previously unseen style (*e.g.*, such as speaker identity and style) derived from a reference utterance. Thanks to the development in TTS, the state-of-the-art (SOTA) style transfer TTS models can realize good style transfer performance with high-quality reference audio. In this paper, we focus on a more challenging setting: the reference utterance contains background noise. Similar to Sytler [14], we conjecture that speaker identity information can be extracted from noisy reference with a noise-robust speaker encoder. However, the style information will be affected by noise [12]. Thus, the problem is to obtain style features from noisy reference similar to those from clean reference.

2.2. Framework overview

We adopt one of the SOTA style transfer TTS models, Gener-Speech [7] as the backbone. The overall architecture of Nore-Speech has been showed in Fig. 2. NoreSpeech is made up of four parts: (1) Encoder, which maps the phoneme sequence into deep representations; (2) DiffStyle, which generates style features based on a noisy spectrogram; (3) Feature fusion, which

combines style and text features; (4) Decoder, which maps the features into mel-spectrogram. The encoder and decoder follow the same structure in the previous work [7].

2.3. DiffStyle

Fig. 2(b) shows the diagram of DiffStyle, which includes three main parts: a conditional diffusion model, a speaker encoder, and two Vector Quantization (VQ) [19] blocks. The conditional diffusion model aims to generate fine-grained style features that represent the speaker's style, and the speaker encoder aims to generate a global speaker embedding that represents the speaker's identity. Both of them take the noisy reference utterance as input. In the following, we will introduce the speaker encoder and conditional diffusion models.

2.3.1. Speaker encoder

We use a generalizable wav2vec 2.0 model [20] to capture the global speaker identity characteristics. Wav2vec 2.0 is a recently proposed self-supervised framework for speech representation learning. In our experiments, we add an average pooling layer and one fully-connected layer on top of the wav2vec 2.0 encoder, which allows us to finetune the encoder on classification tasks. The AMsoftmax [21] loss is employed during finetuning.

2.3.2. Conditional diffusion model

The conditional diffusion model aims to generate noise-agnostic style features based on a noisy audio. To realize this, we adopt the idea of knowledge distillation, which uses a style teacher model to extract style features from clean speech, then the style features are used as the training objective of the diffusion model.

Style teacher models: In this paper, we explore two types of style teacher: (1) Supervised learning (SL) based expressive TTS model, GenerSpeech [7], which can effectively extract style features from clean speech. (2) Self-supervised learning (SSL) [22, 23] based speech decomposition model, NANSY [23], which can extract linguistic and style features from clean speech. We pre-train GenerSpeech and NANSY models in advance, then we take these two kinds of style teacher models to

guide the training of NoreSpeech, denoting as NoreSpeech (T-SL) and NoreSpeech (T-SSL), respectively.

Diffusion model: Diffusion probabilistic (diffusion for short) models [24] have been proved as a powerful generation model in several important domains, *e.g.* image [25], speech [26] and audio [27, 28] fields. The basic idea of diffusion model is to train a neural network for reversing a diffusion process. Given i.i.d. samples $\{\boldsymbol{x}_0 \in \mathbb{R}^D\}$ from an unknown data distribution $p_{data}(\boldsymbol{x}_0)$, diffusion models try to approximate $p_{data}(\boldsymbol{x}_0)$ by a marginal distribution $p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{T-1}|\boldsymbol{x}_T) \cdot p(\boldsymbol{x}_T)dx_{1:T}$.

To implement our conditional diffusion model, we adopt the idea of conditional speech enhancement [18], which uses a shallow convolution layer $\tau_\theta()$ to reshape the noisy mel-spectrogram, then feeds it into a WaveNet-structure diffusion model. In our study, $\boldsymbol{x}_0$ represents style features rather than waveform. The training loss function can be defined as

$$\mathcal{L}_{diff} = \mathbb{E}_{ST(\boldsymbol{y}_c), \boldsymbol{y}_n, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \tau_\theta(\boldsymbol{y}_n))||_2^2 \right] \tag{1}$$

where ST denotes that style teacher model. $\boldsymbol{y}_c$ denotes clean mel-spectrogram, $\boldsymbol{y}_n$ denotes the noisy mel-spectrogram. t is the index of time step. $\boldsymbol{\epsilon}_\theta$ denotes the learnable parameters.

2.3.3. Vector Quantization

Considering the variability of generated style features, we use a Vector Quantization block [19] to map the generated style features into a controllable latent space. We define a latent embedding space $\boldsymbol{e} \in \mathbb{R}^{K \times H}$ where K is the size of the discrete latent space, and H is the dimensionality of each latent embedding vector $\boldsymbol{e}_i$. In our experiments, we set $K = H = 256$. To make sure that the representation sequence commits to an embedding and its output does not grow, a commitment loss is used:

$$\mathcal{L}_c = ||z_e(\boldsymbol{x}) - sg[\boldsymbol{e}]||_2^2 \tag{2}$$

where $z_e(\boldsymbol{x})$ is the output of the vector quantization block, and $sg[\cdot]$ stands for the stop gradient operator.

2.4. Feature fusion

The feature fusion module aims to fuse the phoneme representation and style features. Considering the dimension mismatch between fine-grained style features and the output of the text encoder, we design a parameter-free style-align module to solve this problem. Assume that the time dimensions of style features and text features are t_{style} and t_{text}, respectively. When $t_{style} < t_{text}$, we directly adopt a linear interpolation operation to upsample the style features. When $t_{style} > t_{text}$, we first calculate the ratio between t_{style} and t_{text}, and then we average consecutive frames of style features based on the ratio to downsample the style features.

2.5. Pre-training and loss function

Speaker encoder pre-training: As section 2.3.1 described, we fine-tune the wav2vec 2.0 encoder on LibriTTS dataset, we implement this based on s3prl framework. [1].

Pre-training style teacher: For GenerSpeech teacher, we reproduce GenerSpeech following their paper [7]. The only difference is that we do not use emotion embedding. We train GenerSpeech on the LibriTTS dataset [29]. After that, we use the style adaptor of Generspeech to extract fine-grained prosodic

features from clean speech. For NANSY teacher [23], we first train NANSY [2] on LibriTTS dataset. Then, we use the pretrained model to extract style features. Note that NANSY can extract linguistic, pitch and energy information from the speech, we concatenate the pitch and energy information as the target style features.

Loss function: Our NoreSpeech model can be trained in a end-to-end manner. The training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{dur} + \mathcal{L}_{mel} + \mathcal{L}_{post} + \mathcal{L}_c + \mathcal{L}_{diff} \tag{3}$$

where $\mathcal{L}_{dur}$ denotes the duration prediction loss. $\mathcal{L}_{mel}$ is the mel-spectrogram reconstruction loss. $\mathcal{L}_{post}$ denotes the negative log-likelihood of the post-net [7].

3. Experiment

3.1. Dataset, training setting and baseline models

We train NoreSpeech on LibriTTS-clean dataset [29]. To simulate noisy environments, we use the background sound from the acoustic scene classification task of DCASE 2019 Challenge [30]. All utterances of the noisy speech are mixed with noise sampled from DCASE with an SNR randomly chosen from 5 dB to 25 dB. To evaluate NoreSpeech, we randomly choose 20 sentences test data from LibriTTS test set, which does not appear on the training stage. We conduct preprocessing on the speech data: 1) converting the sampling rate of all data to 16kHz; 2) extracting the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples; 3) converting it into a mel-spectrogram with 80 frequency bins. We train NoreSpeech for 200,000 steps. In the first 50,000 steps, we directly feed the outputs of the style teacher to the feature fusion module. After that, we use the generated style features by the diffusion model as input. For the DiffStyle, the cosine schedule strategy $\beta_t = cos(0.5\pi \cdot \frac{(t/T+s)}{1+s})^2$ is used for any step t, where $s = 0.008$ and $T = 100$. We utilize HiFiGAN [31] as the vocoder to synthesize waveforms from the generated mel-spectrogram.

3.2. Evaluation metrics and Baseline models

For subjective evaluation, we conduct crowd-sourced human evaluations with MOS (mean opinion score) for naturalness and SMOS (similarity mean opinion score) [11] for style similarity on Amazon Mechanical Turk. For objective metrics, we adopt Mel-cepstral distorion (MCD) [32] and Short-Time Objective Intelligibility (STOI) [33] to evaluate the speech quality, and F0 Frame Error (FFE) [34] to evaluate the style similarity.

Baseline models: We compare the quality and similarity of generated audio samples of our NoreSpeech with other systems, including 1) Ground Truth (GT) audio; 2) GT (Mel+HiFiGAN), which means we convert the GT audio into mel-spectrograms and then convert them back to audio using HiFiGAN; 3) Expressive FastSpeech2 [2], which uses the speaker encoder to extract speaker embedding. For a fair comparison, we adopt the same speaker encoder in NoreSpeech for FS2. 4) Styler [14], which uses adversarial training and information bottleneck to eliminate noise; 5) GenerSpeech [7], one of the SOTA style transfer TTS models.

3.3. Experimental results

Table 1 reports the subjective and objective metrics comparison between NoreSpeech and other baseline models. we have

---

[1]https://github.com/s3prl/s3prl

[2]https://github.com/dhchoi99/NANSY

Table 1: *Quality and style similarity results of style transfer. MCD, STIO and FFE are adopted as objective metrics. MOS and SMOS, as subjective metrics, are presented with 95% confidence intervals. For the denoised audio, we one of the SOTA SE models [17] to denoise the noisy reference.*

Method	MCD↓	STIO↑	FFE↓	MOS↑	SMOS↑
GT	-	-	-	4.35 ± 0.09	-
GT(Mel+HiFi-GAN)	4.33	0.971	0.058	4.32 ± 0.09	4.31 ± 0.09
FastSpeech2 (Reference is clean audio) [2]	5.43	0.620	0.28	3.80 ± 0.09	3.86 ± 0.09
FastSpeech2 (Reference is noisy audio)	5.56	0.583	0.308	3.73 ± 0.12	3.72 ± 0.11
Styler (Reference is noisy audio) [14]	5.37	0.655	0.308	3.86 ± 0.11	3.89 ± 0.11
GenerSpeech (Reference is clean audio) [7]	5.29	0.66	0.250	3.93 ± 0.11	4.09 ± 0.11
GenerSpeech (Reference is noisy audio)	5.45	0.619	0.304	3.87 ± 0.12	3.81 ± 0.13
GenerSpeech (Reference is denoised audio)	5.36	0.602	0.257	3.89 ± 0.11	3.95 ± 0.12
NoreSpeech (T-SL) (Reference is noisy audio)	5.25	**0.678**	0.242	3.99 ± 0.10	4.06 ± 0.11
NoreSpeech (T-SSL) (Reference is noisy audio)	**5.02**	0.662	**0.209**	**4.11 ± 0.09**	**4.14 ± 0.09**

Table 2: *The AXY Preference test results. If 7-point score > 0 denotes NoreSpeech has better performance. Preference is calculated based on 7-point score, where 7-point score = 0 denotes raters think the performance of NoreSpeech and baselise is "about the same".*

Baseline	Parallel Style Transfer				Non-Parallel Style Transfer			
	7-point score	Preference (%)			7-point score	Preference (%)		
		Baseline	Same	NoreSpeech		Baseline	Same	NoreSpeech
FS2	1.07	26%	30%	44%	1.48	35%	17%	48%
Styler	1.30	25%	21%	54%	1.26	29%	21%	50%
GenerSpeech	1.20	29%	27%	44%	1.58	26%	12%	62%

Table 3: *Audio quality and similarity comparisons for ablation study. CMOS denotes comparative mean opinion score, CSMOS denotes comparative similarity mean opinion score.*

Setting	CMOS	CSMOS
NoreSpeech	0.0	0.0
w/o VQ	-0.07	-0.12
w/o Style-Align	-0.01	-0.04
w/o DiffStyle	-0.24	-0.32

the following observations: (1) Noise has a significant impact on style transfer performance, *e.g.* the SMOS of GenerSpeech drops from 4.09 to 3.81 when adding noise into reference audio. (2) Comparing to previous SOTA expressive TTS models (Styler and GenerSpeech), our NoreSpeech has better style transfer ability on noisy environment in both subjective and objective evaluations. (3) By comparing NoreSpeech (T-SL) and NoreSpeech (T-SSL), we can see that using an unsupervised speech decomposition (NANSY) as a teacher can bring better performance than using GenerSpeech as the teacher model. We conjecture that NANSY model can extract more robust style features from reference audio due to its self-supervised training strategy. We believe that better style teacher model can be explored to improve the performance of NoreSpeech.

To further evaluate NoreSpeech's style transfer ability, an AXY test [7] of style similarity is conducted to assess the style transfer performance, where raters are asked to rate a 7-point score (from -3 to 3) and choose the speech samples that sound closer to the target style in terms of style expression. We conduct parallel and non-parallel style transfer.

Parallel style transfer (PST): PST denotes that the text input is the same as the reference's content, Table 2 presents the results. Compared to FS2, Styler and Generspeech, our NoreSpeech has better style transfer performance.

Non-parallel style transfer (N-PST): We also explore the ro-

bustness of our NoreSpeech in N-PST, in which a TTS system synthesizes different text in the prosodic style of a reference signal. We can observe that our NoreSpeech significantly improves the model to inform the speaking style, allowing a noisy reference sample to guide the robust stylistic synthesis of arbitrary text. This validates the effectiveness of the straight-forward text-style alignment module in NoreSpeech.

3.4. Ablation study

We have conducted the ablation study to show the effectiveness of each components in NoreSpeech, as shown in Table 3. In our ablation study: (1) To validate the effectiveness of the proposed style-align module, we replace it with the Mel Calibrator [14], which is also parameter-free and has been proved as a better fusion strategy than attention mechanism [35]; (2) To validate our proposed Diffstyle module, we replace it with the multi-level style adaptor [7], which extracts multi-level style features. We have following observations: (1) Our proposed Style-Align module is more effective than Mel Calibrator [14]. (2) VQ module is an important part in our proposed method. (3) DiffStyle module significantly influence the style transfer performance, which shows the effectivess of our proposed Diffstyle.

4. Conclusions

In this paper, we proposed a noise-robust expressive TTS model, named NoreSpeech. Benefitting from DiffStyle and style-align modules, NoreSpeech presents robust stylistic synthesis of arbitrary text, even if the reference audio includes noise. We proved that DiffStyle can be trained with two types of style teacher model, which shows DiffStyle can be further improved through training a better teacher model. We believe DiffStyle can also be used for other tasks (*e.g.* image style transfer). In the future, we will explore better style teacher models and reduce the sample step in DiffStyle.

5. References

[1] J. Shen, R. Pang *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 4779–4783.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[3] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, "Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[4] X. Cheng, Q. Dong, F. Yue, T. Ko, M. Wang, and Y. Zou, "M 3 st: Mix at three levels for speech translation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu *et al.*, "Audiogpt: Understanding and generating speech, music, sound, and talking head," *arXiv preprint arXiv:2304.12995*, 2023.

[6] X. Cheng, Z. Zhu, H. Li, Y. Li, and Y. Zou, "Ssvmr: Saliency-based self-training for video-music retrieval," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis," *arXiv preprint arXiv:2205.07211*, 2022.

[8] Y. Wang, D. Stanton, Y. Zhang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[9] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," *arXiv preprint arXiv:2104.03521*, 2021.

[10] D. Yang, S. Liu, R. Huang, G. Lei, C. Weng, H. Meng, and D. Yu, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *arXiv preprint arXiv:2301.13662*, 2023.

[11] D. Min, D. Lee, E. Yang, and S. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[12] L. Zhang, Y. Ren *et al.*, "Hifidenoise: High-fidelity denoising text to speech with adversarial networks," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7232–7236.

[13] N. Goswami and T. Harada, "SATTS: Speaker attractor text to speech, learning to speak by learning to separate," *arXiv preprint arXiv:2207.06011*, 2022.

[14] K. Lee, K. Park, and D. Kim, "Styler: Style factor modeling with rapidity and robustness via speech decomposition for expressive and controllable neural text to speech," *arXiv preprint arXiv:2103.09474*, 2021.

[15] W. Hsu, Y. Zhang *et al.*, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5901–5905.

[16] K. Nikitaras, G. Vamvoukakis, N. Ellinas *et al.*, "Fine-grained noise control for multispeaker speech synthesis," *arXiv preprint arXiv:2204.05070*, 2022.

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

[18] Y. Lu, Z. Wang *et al.*, "Conditional diffusion probabilistic model for speech enhancement," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7402–7406.

[19] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] A. Baevski, Y. Zhou *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[22] K. Qian, Y. Zhang, S. Chang *et al.*, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[23] H. Choi, J. Lee *et al.*, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[24] J. Sohl, E. Weiss *et al.*, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

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

[26] Z. Kong, W. Ping, J. Huang *et al.*, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[27] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[28] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[29] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[31] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[32] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[34] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3969–3972.

[35] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6264–6268.