

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/369477230>

A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI

Preprint · March 2023


CITATIONS

0

READS

6,019


7 authors, including:



Chenshuang Zhang
Korea Advanced Institute of Science and Technology

19 PUBLICATIONS 48 CITATIONS


SEE PROFILE



Chaoning Zhang
Kyung Hee University

91 PUBLICATIONS 831 CITATIONS


SEE PROFILE



Sheng Zheng
Beijing Institute of Technology

7 PUBLICATIONS 2 CITATIONS

SEE PROFILE



Mengchun Zhang
Korea Advanced Institute of Science and Technology

6 PUBLICATIONS 4 CITATIONS

SEE PROFILE

A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI

CHENSHUANG ZHANG, KAIST, South Korea

CHAONING ZHANG*, Kyung Hee University, South Korea

SHENG ZHENG, Beijing Institute of Technology, China

MENGCHUN ZHANG, KAIST, South Korea

MARYAM QAMAR, Kyung Hee University, South Korea

SUNG-HO BAE, Kyung Hee University, South Korea

IN SO KWEON, KAIST, South Korea

Generative AI has demonstrated impressive performance in various fields, among which speech synthesis is an interesting direction. With the diffusion model as the most popular generative model, numerous works have attempted two active tasks: text to speech and speech enhancement. This work conducts a survey on audio diffusion model, which is complementary to existing surveys that either lack the recent progress of diffusion-based speech synthesis or highlight an overall picture of applying diffusion model in multiple fields. Specifically, this work first briefly introduces the background of audio and diffusion model. As for the text-to-speech task, we divide the methods into three categories based on the stage where diffusion model is adopted: acoustic model, vocoder and end-to-end framework. Moreover, we categorize various speech enhancement tasks by either certain signals are removed or added into the input speech. Comparisons of experimental results and discussions are also covered in this survey.

Additional Key Words and Phrases: Survey, Generative AI, AIGC, Diffusion model, Text to speech, Speech enhancement, Speech synthesis

ACM Reference Format:

Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI. 1, 1 (April 2023), 18 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Recently, generative AI has attracted unprecedented attention in both academia and industry area [140]. Aiming at generating content in different modalities [140], typical tasks in generative AI include but are not limited to chatbot (e.g., ChatGPT [138]), text-to-image synthesis [139] and text-to-speech synthesis [31, 140]. It should be noted that the success of the diffusion model in computer vision has inspired numerous works for speech generation. This work

*Correspondence author.

Authors' addresses: Chenshuang Zhang, KAIST, South Korea, zcs15@kaist.ac.kr; Chaoning Zhang, Kyung Hee University, South Korea, chaoningzhang1990@gmail.com; Sheng Zheng, Beijing Institute of Technology, China, zszhx2021@gmail.com; Mengchun Zhang, KAIST, South Korea, zhangmengchun527@gmail.com; Maryam Qamar, Kyung Hee University, South Korea, maryamqamar@khu.ac.kr; Sung-Ho Bae, Kyung Hee University, South Korea, shbae@khu.ac.kr; In So Kweon, KAIST, South Korea, iskweon77@kaist.ac.kr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

conducts a survey on audio diffusion model for the recent progress in speech synthesis with the focus on text-to-speech synthesis and speech enhancement.

Speech enables humans to express their thoughts and communicate with each other accurately and efficiently. Therefore, speech synthesis is an indispensable component in modern AI system. Specifically, the text-to-speech and speech enhancement task are two main active tasks, which generates a speech from a given text and enhances the quality of an existing speech, respectively. The development of text-to-speech task can be roughly divided into three stages [112]: early works (e.g., formant synthesis [40, 41, 95] and concatenative synthesis [29, 69, 75]), statistical parametric speech synthesis (SPSS)-based methods [115, 116, 133], and neural network-based stage [122, 135, 136]. More recently, diffusion model has attracted great attention in multiple fields (e.g., computer vision) and has also been leveraged into the text-to-speech task [28, 31, 32, 127]. Speech enhancement [74] is another active research field in speech area, which generates speech with a speech signal as input. Common speech enhancement tasks include speech denoising [62], dereverberation [89, 124], and speech super-resolution [46, 56].

Related survey works. Multiple works have conducted a survey on diffusion models covering all fields [6, 129]. Recently, there are also some field-specific survey works, including text-to-image diffusion models [139], graph diffusion models [142]. Complementary to the above works, this work conducts a survey on audio diffusion models. From the perspective of AI-generated content (AIGC), this survey is also related to generative AI (see [140] for a survey) and ChatGPT (see [138] for a survey). From the perspective of speech synthesis, our work is related to multiple survey works [73, 74, 111, 112]. While [111] reviews text-to-speech works, it mainly highlights the digital signal processing components. Other works [73, 74, 112] review speech synthesis based on deep learning, which give us an understanding of the development of neural text-to-speech or speech enhancement. However, as far as we know, our survey is the first work that focuses on reviewing the recent progress in diffusion-based speech synthesis. Specifically, we first briefly introduce the background of audio and speech in section 2. Section 3 and section 4 discusses recent diffusion-based progresses on text-to-speech synthesis and speech enhancement, respectively. Section 5 further summarizes the entire paper.

2 BACKGROUND

2.1 Background on audio and speech

An audio wave in the air is the vibration of its molecules, through which the sound travels. An audio waveform represents the vibration displacement over time, and its strength is indicated by the amplitude [117]. In essence, an audio waveform is a mixture of frequencies [14, 120], and therefore, audio analysis often starts with transforming the audio from its raw waveform in the time domain to the spectrogram in the time-frequency domain. This is achieved by segmenting the audio into windows, for which a short-time Fourier Transform (STFT) [64] calculate its magnitude for each frequency. The STFT is repeated for every window along the time direction, resulting in a map of complex values with two dimensions: X-axis representing frame (time) and Y-axis representing frequency. The complex values on the map can be further converted into their absolute values (or amplitude) and phase. Since human are not sensitive to the frequency equally, a mel-scale is often adopted to transform the spectrogram (magnitude map) into Mel-spectrogram [58, 67]. Mel-spectrogram is widely applied in numerous speech-related tasks [31, 60, 85, 127], including text-to-speech and speech enhancement discussed in this work.

2.2 Background on diffusion model

Before introducing the application of diffusion model in speech synthesis, we first briefly revisit how diffusion model works. As one of the generative model, diffusion model can be traced back to Deep Diffusion Model (DPM) [101] inspired by non-equilibrium thermodynamics. The process of this model is gradually adding noise to a prior distribution and then reversing it to get synthetic data. Denoising diffusion Probabilistic models (DDPM) [24] improved DPM [101] and obtained high-resolution images, which attracted a lot of attention. There is another branch of diffusion models similar to DDPM, the score-based generative model [103, 104]. Although formulated from different perspectives, DDPM [24] and score-based generative model [103, 104] turn out to be equivalent in certain settings. Before we introduce how diffusion model is applied in speech synthesis field, we take DDPM as an example and revisit how DDPM works as follows.

DDPM [24] consists of a forward process and a reverse process. During the forward process, DDPM adds noise to a clean image step by step until the image is destroyed. Define the data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and the latent variable \mathbf{x}_t at step t , the approximate posterior is fixed to a Markov chain and defined as follows (quoted from [24]):

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

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

In the reverse process, DDPM aims to recover the \mathbf{x}_0 from destroyed \mathbf{x}_T . This is trained by estimating the noise during training.

3 TEXT-TO-SPEECH SYNTHESIS

Text-to-speech, also known as speech synthesis, generates speech from text, which has been used in numerous applications.

3.1 Overview of the text-to-speech development

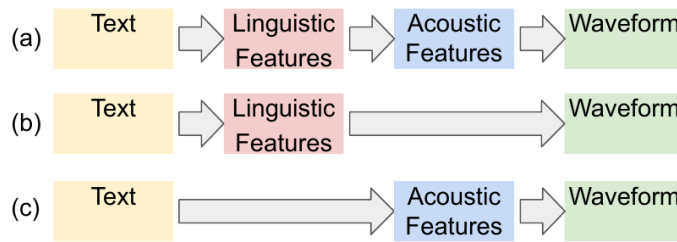


Fig. 1. Development of text-to-speech frameworks. (a) Three-stage framework (b) One branch of two-stage framework that generates waveform from linguistic features directly (c) Another branch of two-stage framework that generates acoustic features from text directly.

From three-stage to two-stage framework. The development of text-to-speech has undergone through a shift from a three-stage framework to a two-stage framework, as shown in Figure 1. Before applying neural networks, statistical

Table 1. Recent progress of text-to-speech diffusion models

Stage	Category	Methods	Resources
Acoustic model	Pioneering work	Diff-TTS [31] Grad-TTS [85]	Code / Project
	Efficient acoustic model	ProDiff [28] DiffGAN-TTS [60]	Project Code
	Adaptive multi-speaker model	Grad-TTS with ILVR [52] Grad-StyleSpeech [32]	Project Project
		Guided-TTS [37] Guided-TTS 2 [38]	Project
	With discrete latent space	Diffsound [128] NoreSpeech [127]	Project
	Fine-grained control	EmoDiff [22]	Project
Vocoder	Pioneering work	WaveGrad [7] DiffWave [45]	Code / Project Code
	Efficient vocoder	BDDM [48] InferGrad [9] WaveFit [43]	Code Project
		DDGM [70] PriorGrad [50] ItôWave [125] SpecGrad [44]	Project Project
	Statistical improvement		
End-to-end	Pioneering work	WaveGrad 2 [8] CRASH [90]	Code / Project Project
	Efficient model	FastDiff [26]	Code / Project
	Further improvements	DAG [79] Itôn [99]	Project

parametric speech synthesis (SPSS) was a popular method [115, 116, 132, 133, 137] consisting of three stages. As shown in Figure 1 (a), the text input is first converted to linguistic features, then acoustic features, and to the waveform in the last stage. Common acoustic features include mel-cepstral coefficients [18], Mel-generalized coefficients [114], F0 [34] and band aperiodicity [33]. Neural networks have brought a paradigm shift from three stages (Figure 1 (a)) to two stages (Figure 1 (b) and Figure 1 (c)). One branch of two-stage framework makes a deep vocoder directly generates waveform from linguistic features (Figure 1 (b)), such as WaveNet [118], Parallel Wavenet [76], DeepVoice 1 [2] DeepVoice 2 [20], HiFi-GAN [4]. Currently, another two-stage framework is more dominant as shown in Figure 1 (c), which directly generates Mel-Spectrogram form of acoustic features from text with a single deep acoustic model, including DeepVoice 3 [83], TransformerTTS [54], Fast Speech 1 [88] and Speech 2 [87].

Overview of diffusion-based methods. Most recent text-to-speech works on diffusion model follow the two-stage framework in Figure 1 (c), which first generate acoustic features with a acoustic models, and then output waveform with a vocoder. Another branch of work attempts to solve the text-to-speech task in an end-to-end manner. We summarize recent work in Table 1, and will introduce each stage within the framework, respectively.

3.2 Acoustic model

Acoustic model that transforms a text to acoustic features is a core component in the the task of text to speech. A summary of representative work that apply diffusion model to acoustic model is shown in Table 2.

Table 2. Experimental results of acoustic model based on diffusion model

Methods	Stage	Dataset	MOS (\uparrow)	RTF (\downarrow)	SMOS (\uparrow)	CER (\downarrow)
Diff-TTS [31]	Acoustic model	LJSpeech	4.337	0.035	-	-
Grad-TTS [85]	Acoustic model	LJSpeech	4.44	0.012	-	-
ProDiff [28]	Acoustic model	LJSpeech	4.08	0.04	-	-
NoreSpeech [127]	Acoustic model	LibriTTS	4.11	-	4.14	-
Grad-TTS with ILVR [52]	Acoustic model	LibriTTS	3.96	-	-	-
Grad-StyleSpeech [32]	Acoustic model	LibriTTS	4.18	-	3.83	2.79
Guided-TTS 2 [38]	Acoustic model	LibriTTS	4.25	-	3.51	0.8
Grad-StyleSpeech [32]	Acoustic model	VCTK	4.13	-	3.95	2.49
Guided-TTS 2 [38]	Acoustic model	VCTK	4.23	-	3.39	0.81

3.2.1 Pioneering works.

In speech synthesis systems, an acoustic model converts the text into acoustic features (e.g., Mel-spectrogram). Diff-TTS [31] is the first work that applies DDPM to el-spectrogram generation. In Diff-TTS [31], a text encoder first extracts the contextual information, which is then aligned with the length predictor and duration predictor. After that, a decoder with the similar architecture as DiffWave [45] is trained to generate mel-spectrogram with DDPM. Specifically, DDIM is leveraged for accelerated sampling. Different from Diff-TTS [31] with DDPM, Grad-TTS [85] is formulated on stochastic differential equation (SDE) [103] and adopts U-net from WaveGrad [7] as the mel-spectrogram generation decoder. Grad-TTS [85] also provides the possibility of an end-to-end text-to-speech pipeline by replacing mel-spectrogram with waveforms as the decoder output.

3.2.2 Towards efficient acoustic model.

Acceleration with knowledge distillation. According to ProDiff [28], estimating the gradient of data density (gradient-based parameterization) is dominant in previous diffusion models, which requires hundreds or thousands of iterations for high quality generation [28]. In order to reduce the iterations as well as maintain quality, ProDiff [28] adopts the *generator-based parameterization* that directly estimates the clean data. Moreover, ProDiff [28] proposes to reduce the data variance by knowledge distillation that optimizes the student model to mimic an N-step DDIM teacher model with N/2 steps. Experimental results show that ProDiff [28] is the first time to make the diffusion models applicable to interactive, real-world speech synthesis applications at a low computation.

Acceleration with Denoising Diffusion GANs. Prior work [126] attributes the thousands of denoising steps in diffusion models to the fact that they commonly approximate the denoising distribution with Gaussian noise, thus requiring small step size. Inspired by [126] that adopts GAN to model the denoising distribution which enables larger step size and less steps, DiffGAN-TTS [60] applies a pretrained GAN as the acoustic generator for acceleration. To further speed up, DiffGAN-TTS [60] also introduces an active shallow diffusion mechanism that conduct denoising conditioned on the coarse prediction by pretrained GAN. Experimental results show that DiffGAN-TTS [60] can generate high-quality audio with only 1 step.

3.2.3 Adaptive modeling for multi-speaker setting.

Prior works [31, 85] generate Mel-spectrogram with a text-conditioned diffusion model. However, they can be trained only when the transcribed data of the target speaker is provided. By applying iterative latent variable sampling [11] to Grad-TTS [85], Grad-TTS with ILVR [52] proposes to mix the latent variable with the reference voice from a target speaker during inference, leading to zero-shot speaker without any training. Also follows Grad-TTS [85], another work Grad-StyleSpeech [32] encodes the mel-spectrogram of reference speech to a styled vector, which is involved in the

training of diffusion models. There is another branch of adaptive modeling applying large-scale untranscribed data in a multi-stage manner. Guided-TTS [37] proposes a two-stage method that first trains an unconditional DDPM model with the large-scale untranscribed data, and then generates the mel-spectrogram with the guidance of phoneme classifiers and the speaker embedding. By contrast, Guided-TTS 2 [38] applies a speaker-conditional DDPM model instead of the unconditional DDPM model in Guided-TTS [37]. Moreover, Guided-TTS 2 [38] adapts the pretrained diffusion model to target speakers with classifier-free guidance and also finetunes the pretrained diffusion model with a short reference speech of the target speaker directly.

3.2.4 Acoustic models with discrete latent space.

A branch of prior works [30, 61] proposes to improve the generation efficiency of auto-regressive methods by compressing the mel-spectrogram into discrete tokens with Vector Quantized Variational Autoencoder (VQ-VAE) [119]. DiffSound [128] follows this discrete setting but proposes a diffusion-based decoder to generate the tokens in a non-autoregressive manner. To solve the data efficiency problem, DiffSound [128] also builds a new text-audio dataset for pertaining, based on the large open-source dataset Audioset [19]. NoreSpeech [127] also adopts VQ-VAE [119] in the generation of mel-spectrogram. Different from DiffSound [128] working on the discretion of mel-spectrogram, NoreSpeech [127] aims to robustly transfer the speaking style to the speech output even when the reference audio contains noise, and thus focuses on the discretion of style features. Specifically, NoreSpeech [127] applies diffusion model to generate *continuous* fine-grained style features conditioned on the noisy reference audio, and then discretizes these features with VQ-VAE [119]. The ground truth style features for diffusion model during training are generated by a pretrained style teacher (e.g., GenSpeech [27] and NANSY [10]) on clean audio. With the fine-grained style features, as well as global speaker embedding, NoreSpeech [127] generates Mel-spectrogram with the same decoder in [27].

3.2.5 Fine-grained control over audio generation.

Controllable emotional models. Most attempts require additional optimization to calculate the emotion intensity values, e.g., relative attributes rank (RAR) [78]. However, these methods may cause degeneration of audio quality. By contrast, EmoDiff [22] proposes a soft-labeled guidance technique to control the emotion intensity directly. Specifically, EmoDiff [22] first trains an unconditional acoustic model, and then trains an emotional classifier on the diffusion trajectory. During inference, the audio is generated under the guidance of the emotional classifier with soft labels. Experimental results show that EmoDiff [22] achieves high generation quality while controlling the emotion precisely.

3.3 Vocoder

Neural vocoders generate waveform based on acoustic feature, e.g., Mel-spectrogram. In earlier researches on vocoders until 2020, autoregressive models have been popular in audio generation for their high-quality output samples but suffer from low inference speed. Although non-autoregressive methods improve the inference speed significantly by reducing sequential steps, there is still an audio quality gap between non-autoregressive and autoregressive methods. A summary of the recent works applying diffusion models to Vocoder is shown in Table 3.

3.3.1 Pioneering works.

WaveGrad [7] is a pioneering work combing score matching and diffusion models by estimating the gradient of the data log-density, which bridges the audio quality gap between non-autoregressive and autoregressive methods. Specifically, WaveGrad [7] proposes two models variants conditioned on discrete refinement step index and continuous noise level, and find that the continuous variant is more effective and flexible considering the various refinement

Table 3. Experimental results of vocoder based on diffusion model

Methods	Stage	Dataset	MOS (\uparrow)	RTF (\downarrow)	PESQ (\uparrow)	STOI (\uparrow)
WaveGrad [7]	Vocoder	LJSpeech	4.47	-	-	-
DiffWave [45]	Vocoder	LJSpeech	4.44	-	-	-
DDGM [70]	Vocoder	LJSpeech	-	-	3.308	0.969
ItôWave [125]	Vocoder	LJSpeech	4.35	-	-	-
InferGrad [9]	Vocoder	LJSpeech	3.97	-	3.578	0.976
BDDM [48]	Vocoder	LJSpeech	4.48	0.438	3.98	0.987

steps during inference. WaveGrad [7] can generate high quality samples with only six refinement steps. Another work DiffWave [45] is the first model showing a high versatility of waveform generation applications based on diffusion models. In the vocoder task, DiffWave [45] is conditioned on mel-spectrogram, and achieves comparable speech quality to the strong autoregressive methods. DiffWave [45] can also produce realistic voices and consistent word-level pronunciation in unconditional and class-conditional settings.

3.3.2 Towards efficient vocoders.

With a shared noise schedule for training and sampling, DDPM [24] requires thousands of sampling iterations for high-quality generation [48]. This property inspires investigations to speed up DDPM [24] by improving the noise schedule. A branch of work on vocoders applies different noise schedules for training and sampling, including WaveGrad [7] and DiffWave [45]. However, these noise schedules are specifically designed and cannot easily generalize to other settings [48]. Another branch of methods searches for a subsequence of the training schedule along the time axis (e.g., DDIM [102]), while it is still challenging to find a short and effective schedule [48].

Schedule prediction by additional networks. In order to find a shorter noise schedule for sampling, [48] proposes an additional schedule network to predict the schedule directly. Together with the original score network in DDPM [24], the model is termed bilateral denoising diffusion model (BDDM) [48]. With both networks, BDDM [48] converges faster during training with the negligible improvement of computation cost than DDPM [24]. In the vocoder task, BDDM [48] can generate indistinguishable samples from human speech with only seven steps, 143x and 28.6x faster than WaveGrad [7] and DiffWave [45], respectively.

Efficient inference by joint training. To reduce the inference iterations while maintaining the generation quality, InferGrad [9] proposes to incorporate the inference process into training with an additional loss. Specifically, InferGrad [9] is optimized to minimize the gap between ground-truth samples and samples generated under inference schedules with a few iterations. Evaluated on LJSpeech dataset, InferGrad [9] achieves comparable voice quality as WaveGrad [7] with 3x speedup.

3.3.3 Improvement from statistical perspective.

Improvement with noise prior. In DDPM [24], Gaussian noise is used in the diffusion process which enables sample arbitrary states without calculating the previous steps. After pointing out that this valuable property stems from the fact the sum of two Gaussian distributions is still Gaussian distribution, denoising diffusion gamma models (DDGM) [70] claims that Gamma distribution can also meet the adding requirement and benefit DDPM by better fitting the estimated noise than Gaussian noise. Following the noise schedules in WaveGrad [7], DDGM [70] improves the quality of generated audio than WaveGrad [7]. Quality improvement is also observed in the image generation area. Another work [50] points out that the Gaussian noise prior may be insufficient to represent all modes of the samples (e.g., the different voiced and

unvoiced segments), leading to a discrepancy between the real data distribution and the choice of prior and thus the training inefficiency. Thus, PriorGrad [50] proposes to apply an adaptive prior from the data statistics for the efficiency improvement of the conditional diffusion model for speech analysis. Specifically, PriorGrad [50] first computes the mean and variance based on the conditional data, and then maps the computed statistics as the mean and variance of Gaussian prior. With the same mean and variance, the noise is similar to data distribution at the instance-level. Empirical study shows that PriorGrad [50] can generate high quality outputs with significantly accelerated inference, either waveform for vocoder or Mel-spectrogram for acoustic model.

Other improvements. ItôWave [125] is the first to propose a vocoder based on linear Itô SDE. Based on Mel-spectrogram, ItôWave [125] achieves higher MOS with 95% confidence than WaveGrad [7] and DiffWave [45]. SpecGrad [44] proposes to adopt the spectral envelope of diffusion noise to the conditional log-mel spectrum, which improves the sound quality especially for the high-quality bands.

3.4 End-to-end frameworks

Instead of treating acoustic modeling and vocoder modeling as independent processes, a branch of work evolves from partially end-to-end methods to fully end-to-end methods gradually. Resembling the two-stage frameworks, partially end-to-end methods [82, 107] also adopt two models as acoustic model and vocoder, but differentiates by training the two models in a joint manner. By contrast, (fully) end-to-end frameworks adopt a single model to generate waveform from text without acoustic features as explicit representation. A line of fully end-to-end work adopts an adversarial decoder (or GAN), including FastSpeech 2 [87], EATS [15] and EFTS-Wav [65]. Most end-to-end methods still rely on generating mel-spectrogram for text-speech alignment, and a spectrogram-free flow-based method is investigated in Wave-Tacotron [123] by simply maximizing likelihood. A limitation of Wave-Tacotron [123] is that the decoder remains autoregressive making it at a disadvantage compared with non-autoregressive counterparts. In this section, we will discuss the recent progresses of end-to-end methods based on diffusion model.

Pioneering works. In contrast to WaveGrad [7] converting Mel-spectrogram to waveform, WaveGrad 2 [8] adopts an end-to-end manner that takes a phoneme sequence as input and generates the audio directly. Specifically, WaveGrad [7] decoder is integrated at the end of Tacotron 2 [98] encoder for a fully differentiable model. As for the duration alignment, WaveGrad 2 [8] generates duration information with a non-attentive Tacotron [97] and optimizes a duration predictor towards ground-truth duration during training. By adjusting the refinement steps, WaveGrad 2 [8] achieves a trade-off between fidelity and speed. Experimental results show that WaveGrad 2 [8] can generate high-quality audio in an end-to-end manner compared to strong baselines. Controllable Raw audio synthesis with High-resolution (CRASH) [90] is a concurrent work to WaveGrad 2 [8] that proposes an end-to-end model for drum sound synthesis. Based on SDE, CRASH [90] applies a noise-conditioned U-Net to estimate the score function, and introduces a class-mixing sampling to generate 'hybrid' sounds. Experimental results show that CRASH [90] can generate drum sounds in multiple tasks, e.g., interpolations and inpainting.

Generation of fullband audios. While previous work focusing on the generation of band-limited audios due to model constraints, DAG [79] adopts an end-to-end manner to generate full-band audios directly. Based on SDE, DAG [79] introduces an encoder-decoder architecture, which downsamples and upsamples the input sequentially. Experimental results show that DAG [79] improves the quality and diversity over existing label-conditioned methods.

Model based on Itô SDE. Inspired by ItôWave [125], Itôn [99] proposes an end-to-end model for speech synthesis based on Itô SDE. Apart from the encoder-decoder architecture, Itôn [99] introduces a dual-denoiser structure for the

generation of mel-spectrogram and waveform, respectively. Moreover, Itôn [99] adopts a two-stage training strategy that trains the encoder and Mel denoiser in the first stage, and the wave denoiser in the second stage.

4 SPEECH ENHANCEMENT

Table 4. Speech enhancement

Task	Category	Methods	Resources
Enhancement by removing	In time-frequency domain	SGMSE [124] SGMSE+ [89] Unfolded CD [130]	Code Code / Project
	In time domain	DiffuSE [62] CDiffuSE [63]	Code
	Unsupervised restoration	UVD [91] Refiner [93]	
Enhancement by adding	Pioneering work	NU-Wave [49] NU-Wave 2 [23]	Code / Project Project
	Further improvement	Improved sampling [134] Improved Diff Wave [141]	Code / Project
More various tasks and frameworks	Source separation	DiffSep [94]	
	Voice conversion	DiffSVC [59]	
	Unified framework for multiple tasks	CQT-Diff [68] UNIVERSE [96]	

Apart from text-to-speech generation, diffusion models have also been widely used in improving the quality of existing degraded audio. Numerous factors can cause the degradation of audio quality and we divide them into two classes according to the restoration type. The first branch of methods removes perturbations in the original clean audio, e.g., noise and reverb. The second branch restores missing parts or adds the desired part, e.g., audio super-resolution. Numerous deep speech enhancement methods have been investigated and they can be categorized into two classes. A discriminative method minimizes the difference between enhanced and clean speech [13, 16, 17, 42], while generative models are optimized by estimating the distribution of clean signals [51, 80, 86, 106, 108]. Despite a superior result regarding objective metrics, the discriminative class often suffers from sounding unnatural compared with the generative class. Diffusion model falling into the generative class is a promising method for bridging its gap with the discriminative class.

4.1 Enhancement by removing

4.1.1 Audio restoration in the time-frequency domain.

A pure generative work. Although formulated on diffusion models, CDiffuSE [63] is trained to estimate the difference between clean and noisy speech. Therefore, it is pointed out in [124] that CDiffuSE [63] can be considered a discriminative task. To make the method *pure generative* and also avoid any noise prior, SGMSE [124] proposes a method based on stochastic differential equations (SDE) [103, 105]. The score function is used to evaluate the quality of the enhanced signal and guide the optimization of the model. In contrast to DiffuSE [62] and CDiffuSE [63] performing diffusion on the time-domain waveform, SGMSE [124] investigates speech enhancement on the time-frequency short-time Fourier transform (STFT) domain to exploit its rich structures. Beyond amplitude of the complex coefficients, it also works on directly enhancing the phase so that an inverse STFT can be applied without extra need of phase retrieval. Experimental results show that SGMSE [124] maintains more natural structures with fewer artifacts than prior work, and achieves quantitative improvements, e.g., an SI-SAR improvement of 3dB over CDiffuSE [63].

Further improvement of SGMSE [124]. SGMSE+ [89] further extends SGMSE [124] with detailed theoretical analysis and improved model architecture. Specifically, SGMSE+ [89] presents a theoretical review of the underlying score matching objective, and investigates different sampling configurations during inference. Inspired by the image generation, SGMSE+ [89] applies a Noise Conditional Score Network (NCSN++) architecture [105], and adapts to the speech area, achieving significant performance improvement over SGMSE [124] on multiple tasks (e.g., speech enhancement and dereverberation). Compared to discriminative methods, SGMSE+ [89] achieves comparable results while showing higher generalization capability in a different corpus from that of training. SGMSE+ [89] also achieves state-of-the-art performance in the single-channel dereverberation task.

Various forms of degradations. Compared to the default Gaussian noise in DDPM [24], cold diffusion [3] improves the sampling procedure in the computer vision area and shows impressive generalization capability in a broad family of degradations, e.g., blur, masking, and downsampling. Unfolded CD [130] follows the procedures in [3] and adapts cold diffusion to the speech enhancement task. Specifically, [130] follows the sampling process of [3] while modifying the degradation progress by containing the clean sample with an iterative interpolation. As for the training strategy, [130] proposes unfolded training method for better error correction in the degradation and restoration process, improving the model performance and stability. Experimental results show that [130] outperforms other diffusion-based models on the speech enhancement task.

4.1.2 Audio restoration in the time domain.

DiffuSE [62] is a pioneering work to apply diffusion model to speech enhancement, also known as audio denoising. In terms of the basic architecture, DiffuSE is inspired by a diffusion-based waveform generative model DiffWave [45]. With the same goal of generating a clean waveform audio, DiffWave [45] conditions on a clean mel-spectrogram, while DiffuSE [62] relies on noisy audio. A major difference of DiffuSE [62] from DiffWave [45] lies in its proposed *supportive reverse process*. Specifically, without access to clean spectral features, DiffuSE [62] resorts to conditioning on noisy spectral features but pretraining the model with clean spectral features as the conditioner. The supportive reverse process uses a noisy speech Mel-spectral sample instead of isotropic gaussian noise as an initial point for a more efficient clean speech recovery. DiffuSE [62] further adopts the fast sampling algorithm developed in DiffWave [45] to speed up its supportive reverse process. On the VoiceBankDEMAND dataset, DiffuSE [62] achieves performance comparable to other time-domain generative methods. In the formulation of DiffuSE [62], it is assumed that isotropic Gaussian noise is applied in the diffusion as well as its reverse process. However, this assumption does not always hold because the noise characteristics are often non-Gaussian. DiffuSE [62] directly combines the noisy speech signal in the sampling process, causing a noise-type mismatch between diffusion and reverse. To address this problem, CDiffuSE [63] formulates a generalized conditional diffusion model to incorporate the observed noisy data, which facilitates estimating both Gaussian noise and non-Gaussian noise. As a result, CDiffuSE [63] achieves superior performance and demonstrates great generalization when discriminative approaches fail.

4.1.3 Unsupervised restoration.

Unsupervised image restoration with diffusion models. Without requiring the pairs of clean and degraded images, unsupervised image restoration [21, 35, 77, 92] has achieved impressive results, but most of them [21, 77, 92] suffer from the heavy computations of iterations. To solve the data problem, denoising diffusion restoration model (DDRM) [35] is the first general sampling-based restoration method that can efficiently generate high-quality, diverse, yet valid solutions for general content images. Specifically, DDRM [35] samples the posterior distribution with a linear degradation operator.

Unsupervised dereverberation. Dereverberation of *natural reverb* has been widely studied in prior work, while removing *artificial reverb* is still challenging since the higher number of variations make supervised methods cannot generalize to unseen data pairs in the training set. To avoid requiring a large amount of paired samples and tackle various types of reverb, [91] proposes an unsupervised vocal dereverberation (UVD) method based on DDRM [35]. Since DDRM [35] assumes a known linear degradation operator, UVD [91] extends DDRM [35] to music dereverberation task where the reverberation operators are always unknown. Specifically, the operator is initially estimated and adaptively corrected by weighted prediction error (WPE) [71]. The predicted clean (dry) signal from a diffusion model is used for operator correction, and the estimated dry signals are generated after iterations. Experimental results show that UVD [91] outperforms unsupervised and supervised benchmarks in both objective and subjective evaluations.

Two-stage refinement in speech enhancement. Research on speech enhancement has achieved significant improvement in terms of signal-to-noise (SNR) ratio but sometimes degrades the speech quality (e.g., naturalness), leading to the degradation of downstream applications. To remove the distortions of speech enhancement outputs, Refiner [93] applies a diffusion model pretrained on clean speech data to detect the degraded part, and then replaces them with newly generated clean ones in the manner of denoising diffusion restoration models(DDRM) [35]. Experimental results show that Refiner [93] is versatile since it improves speech quality with regard to various speech enhancement methods. Moreover, the Refiner [93] can also be integrated into the speech enhancement model for joint optimization in the future.

Table 5. AUDIO RESTORATION RESULTS

Methods	Training set	Testing set	POLQA \uparrow	PESQ \uparrow	ESTOI \uparrow	SI-SDR[dB] \uparrow	SI-SIR[dB] \uparrow	SI-SAR[dB] \uparrow	CSIG \uparrow	CBAK \uparrow	COVL \uparrow
CDiffuSE [63]	WSJ0-CHiME3	WSJ0-CHiME3	2.77 \pm 0.52	2.15 \pm 0.49	0.80 \pm 0.09	7.3 \pm 1.9	19.5 \pm 1.6	7.8 \pm 1.8	—	—	—
SGMSE [124]	WSJ0-CHiME3	WSJ0-CHiME3	2.98 \pm 0.60	2.28 \pm 0.57	0.86 \pm 0.09	14.8 \pm 4.3	25.4 \pm 5.6	15.3 \pm 4.2	—	—	—
SGMSE+ [89]	WSJ0-CHiME3	WSJ0-CHiME3	3.73 \pm 0.53	2.96 \pm 0.55	0.92 \pm 0.06	18.3 \pm 4.4	31.1 \pm 4.6	18.6 \pm 4.5	—	—	—
CDiffuSE [63]	VBD	WSJ0-CHiME3	2.20 \pm 0.50	1.84 \pm 0.41	0.71 \pm 0.10	3.8 \pm 2.5	21.6 \pm 7.0	4.0 \pm 2.5	—	—	—
SGMSE [124]	VBD	WSJ0-CHiME3	2.31 \pm 0.44	1.62 \pm 0.31	0.74 \pm 0.11	10.5 \pm 4.1	20.2 \pm 7.6	11.4 \pm 3.5	—	—	—
SGMSE+ [89]	VBD	WSJ0-CHiME3	3.43 \pm 0.61	2.48 \pm 0.58	0.90 \pm 0.07	16.2 \pm 4.1	28.9 \pm 4.6	16.4 \pm 4.1	—	—	—
DiffuSE [62]	VBD	VBD	—	2.43	—	10.5 \pm 0.14	30.0 \pm 0.71	10.8 \pm 0.11	3.63	2.81	3.01
CDiffuSE [63]	VBD	VBD	—	2.52	0.79	12.6/12.1 \pm 0.10	28.2 \pm 0.36	12.3 \pm 0.09	3.72	2.91	3.10
SGMSE [124]	VBD	VBD	—	2.28	0.80	16.2/15.1 \pm 0.27	24.9 \pm 0.42	15.7 \pm 0.25	—	—	—
SGMSE+ [89]	VBD	VBD	—	2.93	0.87	17.3	—	—	—	—	—
UVD [91]	NHSS	NHSS	—	—	—	—	—	—	—	—	—
Unfolded CD [130]	VBD	VBD	—	2.77	—	—	—	—	3.91	3.32	3.33
Refiner [93]	VBC	VBC	—	—	—	—	—	—	—	—	—

4.2 Enhancement by adding

4.2.1 Pioneering works on audio super-resolution.

Audio super-resolution [46, 56], also widely known as upsampling [84] or bandwidth extension [46], aims to generate audio of a high sampling rate from that of a low sampling rate via extending its bandwidth. Compared to previous work with 16kHz as the target frequency [5, 25, 39, 46, 53, 55, 56, 84], NU-Wave [49] is the first to synthesize 48kHz waveforms from 16kHz or 24kHz inputs, and also the first to apply diffusion models for audio super-resolution. Following the structures of prior diffusion-based vocoders (DiffWave [45] and WaveGrad [7]), NU-Wave [49] adapts the model to audio super-resolution task, e.g. introducing a sinusoidal 128-dimensional vector as noise level embedding. Moreover, NU-Wave [49] empirically finds that the receptive field for a conditional signal should be larger than that of the noisy input, and thus modifies the local conditioner for the information ensemble of signals with different receptive fields and upscaling ratios. Experimental results show that NU-Wave [49] outperforms the baseline methods in all cases, even with smaller model capacity.

NU-Wave 2 [23] further improves NU-Wave [49] from two aspects. On the one hand, NU-Wave 2 [23] adopts short-time Fourier convolution (STFC) to overcome the limitations of NU-Wave [49] that fails to generate harmonics of vowels [49] and various frequency bands [57, 72]. On the other hand, different from prior models that the initial and target sampling rates are fixed, NU-Wave 2 [23] defines a new task *general neural audio upsampling* that the inputs can be any sampling rate for a single model. Specifically, NU-Wave 2 [23] proposes bandwidth spectral feature transform (BSFT) layer for adapting to various sampling rates. Experimental results show that NU-Wave 2 [23] has many valuable characteristics compared to baseline methods, e.g., only NU-Wave 2 [23] can generate harmonics.

Improved sampling method. Diffusion-based audio super-resolution is commonly conducted by conditioning the denoising network on low-resolution audio. [134] further improves the audio quality by injecting the low-resolution audio into the sampling process as a condition if the downsampling schedule is known. This sampling method can be used directly in other diffusion-based super-resolution methods. Moreover, [134] can generalize to multiple settings with a UDM, e.g., varying upscaling ratios. Experimental results show that [134] achieves the state-of-the-art log-spectral-distance (LSD) on the 48kHz VCTK Multi-Speaker benchmark.

Improved model architecture. [141] improves the speech quality caused by deterministic mathematical degradation, e.g., compression, clipping and downsampling. Conditioned on degraded Mel-spectrogram, [141] finds that DiffWave [45] can restore the degraded speech to some extent. For further improvement, [141] replaces the original upsampler of DiffWave [45] with a CNN sampler and trains the CNN sampler separately with the original upsampler as the reference. We summarize the experimental results of speech enhancement models in Table 5.

4.3 Miscallenous audio tasks

Source separation. Source separation is a task that aims to recover speech of interest from a mixed signal [94].

Early work on source separation is based on short-time Fourier transform (STFT), which aims to get the clean spectrum from the mixed spectrum. However, using STFT in audio separation is not sensitive to phase and needs high-frequency resolution, making it low accuracy of reconstruction and time-consuming. To overcome these problems, other time-domain approaches are proposed, including independent component analysis (ICA) [12], non-negative matrix factorization(NMF) [131].DiffSep [94] is the first diffusion-based method for single-channel speech separation. Specifically, DiffSep [94] trains a SDE model that converts the separated signal to the mixtures, thus the reverse process can separate individual source from the mixed signal. DiffSep [94] can also be applied for audio denoising since noise can be seen as a type of extra source.

Voice conversion. Voice Conversion edits the source speakers to adapt to the target speaker by changing speech signal features. Traditional techniques include vector quantization [1, 100], hidden Markov models [36, 113], and Gaussian mixture models [109, 121]. However, these methods only focus on specific parts of spectrum rather than the entire spectrum, resulting in poor speech quality. To solve these problems, [47, 66, 81, 110?] adopt neural networks which cover the entire spectrum, and DiffSVC [59] is a pioneering work that leveraging diffusion models on voice conversion. Specifically, DiffSVC [59] is developed for singing voice conversion (SVC), which first transforms the phonetic posteriorgrams (PPGs) to spectral features, and then transforms them into waveforms by a trained neural vocoder. This method can get the target speech by any other speech as input (i.e., any-to-one SVC).

Towards a Unified framework for multiple tasks. A branch of work aims to solve multiple tasks with a single model [68, 96]. CQT-Diff [68] explores whether a pretrained model can be used for different tasks without knowing the degradation type during inference. Specifically, CQT-Diff [68] excels at the bandwidth extension task, and achieves competitive results to baseline methods in long audio inpainting and declipping tasks without retraining. Note that the

method is termed as CQT-Diff since Constant-Q Transform (CQT) is adopted for strong harmonic signals (e.g., music). Except common background noise and reverberation, another work [96] points out that a large number of distortions may exist in the audios, including bandwidth reduction, clipping, silent gaps. Therefore, [96] proposes a universal system that can tackle 55 different distortions in an end-to-end manner. With a conditioner network that tackles the main part of a task, a generator network is formulated on the score-based diffusion models for waveform synthesis. Experimental results show that this approach achieves the state-of-the-art performance according to a subjective text with expert listeners [96].

5 CONCLUSION

This survey reviews the recent progress of speech synthesis based on diffusion model. After introducing the background of audio signals and diffusion model, we discuss recent works of two main tasks in the speech synthesis field: text-to-speech and speech enhancement. We hope this survey could provide an insightful understanding for the researchers who are interested in speech synthesis as well as generative AI.

REFERENCES

- [1] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. 1990. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)* 11, 2 (1990), 71–76.
- [2] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*. PMLR, 195–204.
- [3] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392* (2022).
- [4] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. 2019. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646* (2019).
- [5] Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W Koh, and Stefano Ermon. 2019. Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations. *Advances in Neural Information Processing Systems* 32 (2019).
- [6] Hanqun Cao, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2022. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646* (2022).
- [7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. WaveGrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713* (2020).
- [8] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. 2021. WaveGrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660* (2021).
- [9] Zehua Chen, Xu Tan, Ke Wang, Shifeng Pan, Danilo Mandic, Lei He, and Sheng Zhao. 2022. Infergrad: Improving Diffusion Models for Vocoder by Considering Inference in Training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8432–8436.
- [10] Hyeon-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems* 34 (2021), 16251–16265.
- [11] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938* (2021).
- [12] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. 2005. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews* 6, 1 (2005), 1–57.
- [13] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847* (2020).
- [14] SA Dobrynin, EA Kolubaev, A Yu Smolin, AI Dmitriev, and SG Psakhie. 2010. Time-frequency analysis of acoustic signals in the audio-frequency range generated during Hadfield’s steel friction. *Technical Physics Letters* 36 (2010), 606–609.
- [15] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. 2020. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575* (2020).
- [16] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*. PMLR, 2031–2041.
- [17] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai. 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 9 (2018),

- 1570–1584.
- [18] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. 1992. An adaptive algorithm for mel-cepstral analysis of speech.. In *icassp*, Vol. 92. 137–140.
 - [19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.
 - [20] Andrew Gibiansky, Serkan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems* 30 (2017).
 - [21] Jinjin Gu, Yujun Shen, and Bolei Zhou. 2020. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3012–3021.
 - [22] Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2022. EmoDiff: Intensity Controllable Emotional Text-to-Speech with Soft-Label Guidance. *arXiv preprint arXiv:2211.09496* (2022).
 - [23] Seungu Han and Junhyeok Lee. 2022. NU-Wave 2: A general neural audio upsampling model for various sampling rates. *arXiv preprint arXiv:2206.08545* (2022).
 - [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
 - [25] Nana Hou, Chenglin Xu, Joey Tianyi Zhou Van Tung Pham, Joey Tianyi Zhou, Eng Siong Chng, and Haizhou Li. 2020. Speaker and Phoneme-Aware Speech Bandwidth Extension with Residual Dual-Path Network.. In *Interspeech*. 4064–4068.
 - [26] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. *arXiv preprint arXiv:2204.09934* (2022).
 - [27] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022. GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech Synthesis. *arXiv preprint arXiv:2205.07211* (2022).
 - [28] 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. *arXiv preprint arXiv:2207.06389* (2022).
 - [29] Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1. IEEE, 373–376.
 - [30] Vladimir Iashin and Esa Rahtu. 2021. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791* (2021).
 - [31] 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. *arXiv preprint arXiv:2104.01409* (2021).
 - [32] Minki Kang, Dongchan Min, and Sung Ju Hwang. 2022. Any-speaker Adaptive Text-To-Speech Synthesis with Diffusion Models. *arXiv preprint arXiv:2211.09383* (2022).
 - [33] Hideki Kawahara, Jo Estill, and Osamu Fujimura. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Second international workshop on models and analysis of vocal emissions for biomedical applications*.
 - [34] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication* 27, 3-4 (1999), 187–207.
 - [35] Bahjat Kwar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793* (2022).
 - [36] Eun-Kyoung Kim, Sangho Lee, and Yung-Hwan Oh. 1997. Hidden Markov model based voice conversion using dynamic characteristics of speaker. In *European Conference On Speech Communication And Technology*. Eurospeech, 2519–2522.
 - [37] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. 2021. Guided-TTS: Text-to-Speech with Untranscribed Speech. *arXiv preprint arXiv:2111.11755* (2021).
 - [38] Sungwon Kim, Heeseung Kim, and Sungroh Yoon. 2022. Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data. *arXiv preprint arXiv:2205.15370* (2022).
 - [39] Sung Kim and Visvesh Sathe. 2019. Bandwidth extension on raw audio via generative adversarial networks. *arXiv preprint arXiv:1903.09027* (2019).
 - [40] Dennis H Klatt. 1980. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America* 67, 3 (1980), 971–995.
 - [41] Dennis H Klatt. 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82, 3 (1987), 737–793.
 - [42] Yuma Koizumi, Kenta Niwa, Yusuke Hioka, Kazunori Kobayashi, and Yoichi Haneda. 2018. DNN-based source enhancement to increase objective sound quality assessment score. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 10 (2018), 1780–1792.
 - [43] Yuma Koizumi, Kohei Yatabe, Heiga Zen, and Michiel Bacchiani. 2022. WaveFit: An Iterative and Non-autoregressive Neural Vocoder based on Fixed-Point Iteration. *arXiv preprint arXiv:2210.01029* (2022).
 - [44] Yuma Koizumi, Heiga Zen, Kohei Yatabe, Nanxin Chen, and Michiel Bacchiani. 2022. SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping. *arXiv preprint arXiv:2203.16749* (2022).
 - [45] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).

- [46] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. 2017. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853* (2017).
- [47] Jiahao Lai, Bo Chen, Tian Tan, Sibong Tong, and Kai Yu. 2016. Phone-aware LSTM-RNN for voice conversion. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 177–182.
- [48] Max WY Lam, Jun Wang, Dan Su, and Dong Yu. 2022. BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis. *arXiv preprint arXiv:2203.13508* (2022).
- [49] Junhyeok Lee and Seungu Han. 2021. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321* (2021).
- [50] Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. 2021. Priorgrad: Improving conditional denoising diffusion models with data-driven adaptive prior. *arXiv preprint arXiv:2106.06406* (2021).
- [51] Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. 2020. A recurrent variational autoencoder for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 371–375.
- [52] Alon Levkovitch, Eliya Nachmani, and Lior Wolf. 2022. Zero-Shot Voice Conditioning for Denoising Diffusion TTS Models. *arXiv preprint arXiv:2206.02246* (2022).
- [53] Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee. 2015. DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- [54] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6706–6713.
- [55] Xinyu Li, Venkata Chebriyyam, Katrin Kirchhoff, and AI Amazon. 2019. Speech Audio Super-Resolution for Speech Recognition.. In *INTERSPEECH*. 3416–3420.
- [56] Teck Yian Lim, Raymond A Yeh, Yijia Xu, Minh N Do, and Mark Hasegawa-Johnson. 2018. Time-frequency networks for audio super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 646–650.
- [57] Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. 2021. VoiceFixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731* (2021).
- [58] Qingzhong Liu, Andrew H Sung, and Mengyu Qiao. 2009. Temporal derivative-based spectrum and mel-cepstrum audio steganalysis. *IEEE Transactions on Information Forensics and Security* 4, 3 (2009), 359–368.
- [59] Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. 2021. Diffsvc: A diffusion probabilistic model for singing voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 741–748.
- [60] Songxiang Liu, Dan Su, and Dong Yu. 2022. DiffGAN-TTS: High-Fidelity and Efficient Text-to-Speech with Denoising Diffusion GANs. *arXiv preprint arXiv:2201.11972* (2022).
- [61] Kubo Liu, Turab Iqbal, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2021. Conditional sound generation using neural discrete time-frequency representation learning. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.
- [62] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. 2021. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 659–666.
- [63] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. 2022. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7402–7406.
- [64] Paul Magron, Roland Badeau, and Bertrand David. 2018. Model-based STFT phase recovery for audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 6 (2018), 1095–1105.
- [65] Chenfeng Miao, Liang Shuang, Zhengchen Liu, Chen Minchuan, Jun Ma, Shaojun Wang, and Jing Xiao. 2021. Efficienttts: An efficient and high-quality text-to-speech architecture. In *International Conference on Machine Learning*. PMLR, 7700–7709.
- [66] Shariq Mobin and Joan Bruna. 2016. Voice conversion using convolutional neural networks. *arXiv preprint arXiv:1610.08927* (2016).
- [67] Sirko Molau, Michael Pitz, Ralf Schluter, and Hermann Ney. 2001. Computing mel-frequency cepstral coefficients on the power spectrum. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221)*, Vol. 1. IEEE, 73–76.
- [68] Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. 2022. Solving Audio Inverse Problems with a Diffusion Model. *arXiv preprint arXiv:2210.15228* (2022).
- [69] Eric Moulines and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication* 9, 5-6 (1990), 453–467.
- [70] Eliya Nachmani, Robin San Roman, and Lior Wolf. 2021. Denoising diffusion gamma models. *arXiv preprint arXiv:2110.05948* (2021).
- [71] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. 2010. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7 (2010), 1717–1731.
- [72] Viet-Anh Nguyen, Anh HT Nguyen, and Andy WH Khong. 2022. TUNet: A Block-online Bandwidth Extension Model based on Transformers and Self-supervised Pretraining. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 161–165.
- [73] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences* 9, 19 (2019), 4050.

- [74] Ramesh Nuthakki, Payel Masanta, and TN Yukta. 2022. A Literature Survey on Speech Enhancement Based on Deep Neural Network Technique. In *ICCCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering*. Springer, 7–16.
- [75] Joseph Olive. 1977. Rule synthesis of speech from dyadic units. In *ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. IEEE, 568–570.
- [76] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*. PMLR, 3918–3926.
- [77] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. 2021. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [78] Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *2011 International Conference on Computer Vision*. IEEE, 503–510.
- [79] Santiago Pascual, Gautam Bhattacharya, Chunghsin Yeh, Jordi Pons, and Joan Serrà. 2022. Full-band General Audio Synthesis with Score-based Diffusion. *arXiv preprint arXiv:2210.14661* (2022).
- [80] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).
- [81] Santiago Pascual and Antonio Bonafonte Cávez. 2016. Multi-output RNN-LSTM for multiple speaker speech synthesis with a-interpolation model. In *SSW9: 9th ISCA Workshop on Speech Synthesis: proceedings: Sunnyvale (CA, USA): September 13-15, 2016*. Institute of Electrical and Electronics Engineers (IEEE), 112–117.
- [82] Wei Ping, Kainan Peng, and Jitong Chen. 2018. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281* (2018).
- [83] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. (2017).
- [84] Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà. 2021. Upsampling artifacts in neural audio synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3005–3009.
- [85] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. 8599–8608.
- [86] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. 2017. Speech Enhancement Using Bayesian Wavenet. In *Interspeech*. 2013–2017.
- [87] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. FastSpeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).
- [88] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems* 32 (2019).
- [89] Julius Richter, Simon Welker, Jean-Marie Lemerrier, Bunlong Lay, and Timo Gerkmann. 2022. Speech enhancement and dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2208.05830* (2022).
- [90] Simon Rouard and Gaëtan Hadjeres. 2021. CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis. *arXiv preprint arXiv:2106.07431* (2021).
- [91] Koichi Saito, Naoki Murata, Toshimitsu Uesaka, Chieh-Hsin Lai, Yuhta Takida, Takao Fukui, and Yuki Mitsufuji. 2022. Unsupervised vocal dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2211.04124* (2022).
- [92] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems* 32 (2019).
- [93] Ryosuke Sawata, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. 2022. A Versatile Diffusion-based Generative Refiner for Speech Enhancement. *arXiv preprint arXiv:2210.17287* (2022).
- [94] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaekuk Byun, Soyeon Choe, and Min-Seok Choi. 2022. Diffusion-based Generative Speech Source Separation. *arXiv preprint arXiv:2210.17327* (2022).
- [95] P Seeviour, J Holmes, and M Judd. 1976. Automatic generation of control signals for a parallel formant speech synthesizer. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, 690–693.
- [96] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. 2022. Universal Speech Enhancement with Score-based Diffusion. *arXiv preprint arXiv:2206.03065* (2022).
- [97] Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. 2020. Non-attentive tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301* (2020).
- [98] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.
- [99] Ziqiang Shi and Shoule Wu. 2022. ITÖN: End-to-end audio generation with Itô stochastic differential equations. *Digital Signal Processing* 132 (2022), 103781.
- [100] Kiyohiro Shikano, Kai-Fu Lee, and Raj Reddy. 1986. Speaker adaptation through vector quantization. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 11. IEEE, 2643–2646.

- [101] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. 2256–2265.
- [102] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- [103] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution, Vol. 32.
- [104] Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems* 33 (2020), 12438–12448.
- [105] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- [106] Meet H Soni, Neil Shah, and Hemant A Patil. 2018. Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5039–5043.
- [107] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. (2017).
- [108] Martin Strauss and Bernd Edler. 2021. A flow-based neural network for time domain speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5754–5758.
- [109] Yannis Stylianou, Olivier Cappé, and Eric Moulines. 1998. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing* 6, 2 (1998), 131–142.
- [110] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4869–4873.
- [111] Youcef Tabet and Mohamed Boughazi. 2011. Speech synthesis techniques. A survey. In *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*. IEEE, 67–70.
- [112] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561* (2021).
- [113] Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. 1995. Speech parameter generation from HMM using dynamic features. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, 660–663.
- [114] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. 1994. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation.. In *ICSLP*, Vol. 94. 18–22.
- [115] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. 2013. Speech synthesis based on hidden Markov models. *Proc. IEEE* 101, 5 (2013), 1234–1252.
- [116] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, Vol. 3. IEEE, 1315–1318.
- [117] George Tzanetakis and Perry Cook. 2000. Marsyas: A framework for audio analysis. *Organised sound* 4, 3 (2000), 169–175.
- [118] 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*.
- [119] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS* (2017).
- [120] Sean Vasequez and Mike Lewis. 2019. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083* (2019).
- [121] Christophe Veaux and Xavier Rodet. 2011. Intonation conversion from neutral to expressive speech. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [122] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).
- [123] Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P Kingma. 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5679–5683.
- [124] Simon Welker, Julius Richter, and Timo Gerkmann. 2022. Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain. *arXiv preprint arXiv:2203.17004* (2022).
- [125] Shoule Wu and Ziqiang Shi. 2022. ItôWave: Itô Stochastic Differential Equation is all You Need for Wave Generation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8422–8426.
- [126] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. 2021. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804* (2021).
- [127] Dongchao Yang, Songxiang Liu, Jianwei Yu, Helin Wang, Chao Weng, and Yuexian Zou. 2022. NoreSpeech: Knowledge Distillation based Conditional Diffusion Model for Noise-robust Expressive TTS. *arXiv preprint arXiv:2211.02448* (2022).
- [128] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *arXiv preprint arXiv:2207.09983* (2022).
- [129] Ruihan Yang, Prakhara Srivastava, and Stephan Mandt. 2022. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481* (2022).
- [130] Hao Yen, François G Germain, Gordon Wichern, and Jonathan Le Roux. 2022. Cold Diffusion for Speech Enhancement. *arXiv preprint arXiv:2211.02527* (2022).

- [131] Kazuyoshi Yoshii, Ryota Tomioka, Daichi Mochihashi, and Masataka Goto. 2013. Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction. In *ISMIR*. 369–374.
- [132] Takayoshi Yoshimura. 2002. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. *PhD diss, Nagoya Institute of Technology* (2002).
- [133] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*.
- [134] Chin-Yun Yu, Sung-Lin Yeh, György Fazekas, and Hao Tang. 2022. Conditioning and Sampling in Variational Diffusion Models for Speech Super-resolution. *arXiv preprint arXiv:2210.15793* (2022).
- [135] Heiga Ze, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7962–7966.
- [136] Heiga Zen and Haşim Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4470–4474.
- [137] Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. *speech communication* 51, 11 (2009), 1039–1064.
- [138] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Sheng Zheng, Yu Qiao, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Gyeong-Moon Park, Jinwoo Choi, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023. One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *researchgate DOI:10.13140/RG.2.2.24789.70883* (2023).
- [139] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image Diffusion Model in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909* (2023).
- [140] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *arXiv preprint arXiv:2303.11717* (2023).
- [141] Jianwei Zhang, Suren Jayasuriya, and Visar Berisha. 2021. Restoring degraded speech via a modified diffusion model. *arXiv preprint arXiv:2104.11347* (2021).
- [142] Mengchun Zhang, Maryam Qamar, Taegoo Kang, Yuna Jung, Chenshuang Zhang, Sung-Ho Bae, and Chaoning Zhang. 2023. A Survey on Graph Diffusion Models: Generative AI in Science for Molecule, Protein and Material. *ResearchGate 10.13140/RG.2.2.26493.64480* (2023).