Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey

Tianxin Xie*, Yan Rong*, Pengfei Zhang*, Wenwu Wang, Li Liu

***Abstract*—Text-to-speech (TTS), also known as speech synthesis, is a prominent research area that aims to generate natural-sounding human speech from text. Recently, with the increasing industrial demand, TTS technologies have evolved beyond synthesizing human-like speech to enabling controllable speech generation. This includes fine-grained control over various attributes of synthesized speech such as emotion, prosody, timbre, and duration. In addition, advancements in deep learning, such as diffusion and large language models, have significantly enhanced controllable TTS over the past several years. In this work, we conduct a comprehensive survey of controllable TTS, covering approaches ranging from basic control techniques to methods utilizing natural language prompts, aiming to provide a clear understanding of the current state of research. We examine the general controllable TTS pipeline, challenges, model architectures, and control strategies, offering a comprehensive and clear taxonomy of existing methods. Additionally, we provide a detailed summary of datasets and evaluation metrics and shed some light on the applications and future directions of controllable TTS. To the best of our knowledge, this survey paper provides the first comprehensive review of emerging controllable TTS methods, which can serve as a beneficial resource for both academic researchers and industrial practitioners.**

***Index Terms*—Text-to-speech, controllable TTS, speech synthesis, TTS survey, large language models, diffusion models.**

I. Introduction

Speech synthesis, also broadly known as text-to-speech (TTS), is a long-time developed technique that aims to synthesize human-like voices from text [1], [2], and it has extensive applications in our daily lives, such as health care [3], [4], personal assistants [5], entertainment [6], [7], and robotics [8], [9]. Recently, TTS has gained significant attention with the rise of large language model (LLM)-powered chatbots, such as ChatGPT [10] and LLaMA [11], due to its naturalness and convenience for human-computer interaction. Meanwhile, the ability to achieve fine-grained control over synthesized speech attributes, such as emotion, prosody, timbre, and duration, has become a hot research topic in both academia and industry, driven by its vast potential for diverse applications.

Deep learning [12] has made great progress in the past decade due to exponentially growing computational resources like GPUs [13], leading to the explosion of numerous exciting works on TTS [14]–[17]. These methods can synthesize human speech with improved quality [14] and can achieve fine-grained control of the generated voice [18]–[22]. In addition, some recent works synthesize speech given multimodal input, such as face images [23], [24], cartoons [7], and videos [25]. Moreover, with the fast development of open-source LLMs [11], [26]–[29], some researchers propose to synthesize fine-grained controllable speech with natural language description [30]–[32], offering a new way to generate custom speech voices. Meanwhile, powering LLMs with speech synthesis has also been a hot topic in the last few years [33]–[35]. In recent years, a wide range of TTS methods has emerged, making it essential for researchers to gain a comprehensive understanding of current research trends, particularly in controllable TTS, and to identify promising future directions in this rapidly evolving field. Consequently, there is a pressing need for an up-to-date survey of TTS techniques. While several existing surveys address parametric approaches [36]–[41] and deep learning-based approaches [42]–[48], they largely overlook the controllability of TTS. Additionally, these surveys do not cover recent advancements, such as natural language description-based TTS methods.

This paper provides a comprehensive and in-depth survey of existing and emerging TTS technologies, with a particular focus on controllable TTS methods. Fig. 1 demonstrates the development of controllable TTS methods in recent years, showing their backbones, feature representations, and control abilities. The remainder of this section begins with a brief comparison between this survey and previous ones, followed by an overview of the history of controllable TTS technologies, ranging from early milestones to state-of-the-art advancements. Finally, we introduce the taxonomy and organization of this paper. We have posted a version of our paper on arXiv.org (https://arxiv.org/abs/2412.06602).

A. Comparison with Existing Surveys

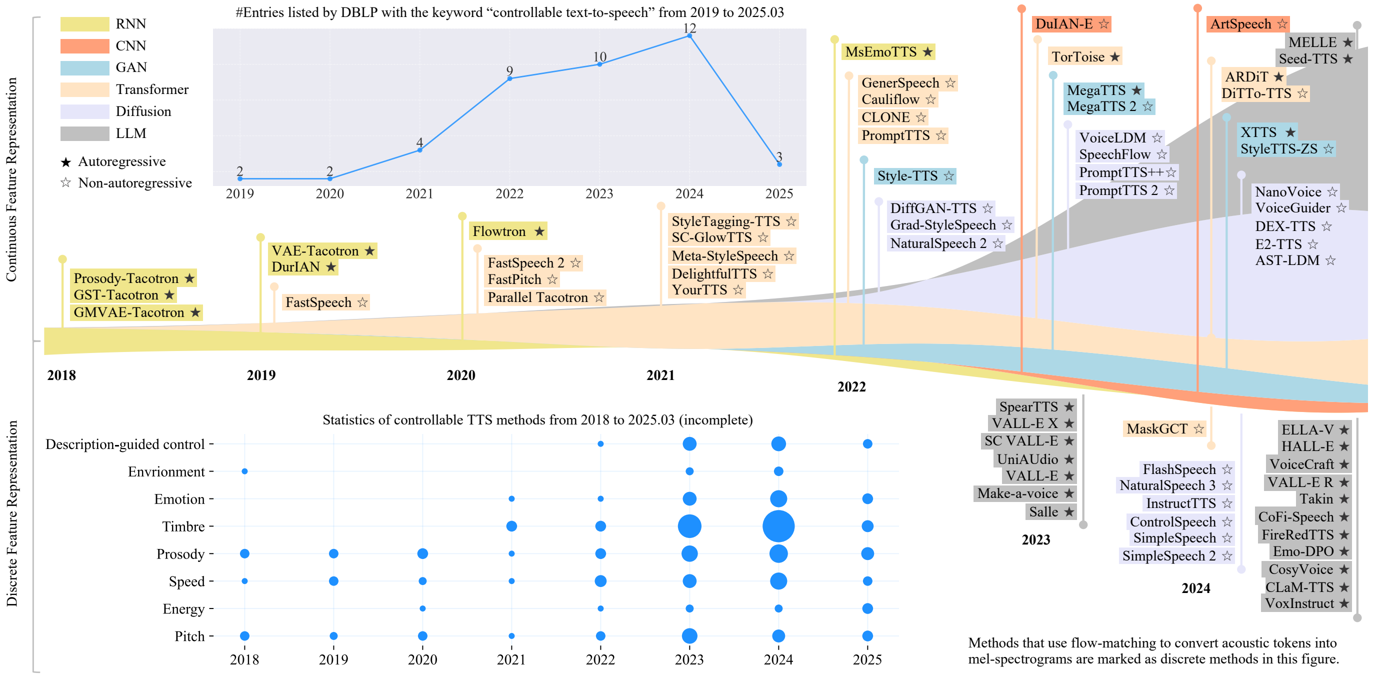
Several survey papers have reviewed TTS technologies, spanning early approaches from previous decades [36], [37], [40], [49] to more recent advancements [42], [43], [50]. However, to the best of our knowledge, this paper is the first to focus specifically on controllable TTS. The key differences between this survey and prior work are summarized as follows:

Different Scope. Klatt et al. [36] provided the first comprehensive survey on formant, concatenative, and articulatory TTS methods, with a strong emphasis on text analysis. In the early 2010s, Tabet et al. [49] and King et al. [40] explored rule-based, concatenative, and Hidden Markov Models (HMM)-based techniques. Later, the advent of deep learning catalyzed the emergence of numerous neural model-based TTS methods.

Tianxin Xie, Yan Rong, Pengfei Zhang and Li Liu are with the Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511458, China. Wenwu Wang is with Surrey University, UK. Corresponding author: Li Liu, avrillliu@hkust-gz.edu.cn.

* Equal contribution.

Readers can check this GitHub repository (https://github.com/imxtx/awesome-controllabe-speech-synthesis) for updates and discussion.



Methods that use flow-matching to convert acoustic tokens into mel-spectrograms are marked as discrete methods in this figure.

Fig. 1. A summary of representative controllable TTS methods in recent years and their model architectures, feature representations, and control abilities. Additional network structures, such as VAE and flow-based models, are not included in this figure. For more details, refer to Tables IV and III.

Therefore, Ning et al. [43] and Tan et al. [42] have conducted extensive surveys on neural acoustic models and vocoders, while Zhang et al. [50] presented the first review of diffusion model-based TTS techniques. However, these studies offer limited discussion on the controllability of TTS systems. To address this gap, we present the first comprehensive survey of TTS methods through the lens of controllability, providing an in-depth analysis of model architectures and strategies for controlling synthesized speech.

Close to Current Demands. With the rapid development of hardware (i.e., GPUs and TPUs) and artificial intelligence (AI) techniques (i.e., transformers, LLMs, diffusion models) in the last few years, the demand for controllable TTS is becoming increasingly urgent due to its broad applications in industries such as filmmaking, gaming, robotics, and personal assistants. Despite this growing need, existing surveys pay little attention to control methods in TTS technologies. To bridge this gap, we propose a systematic analysis of current controllable TTS methods and the associated challenges, offering a comprehensive understanding of the research state in this field.

New Insights and Directions. This survey offers new insights through a comprehensive analysis of model architectures and control methods in controllable TTS systems. Additionally, it provides an in-depth discussion of the challenges associated with various controllable TTS tasks. Furthermore, we address the question: "Where are we on the path to fully controllable TTS technologies?", by examining the relationship and gap between current TTS methods and industrial requirements. Based on these analyses, we identify promising directions for future research on TTS technologies.

Table I summarizes the existing surveys and our survey in terms of main focus and publication year.

TABLE I
COMPARISON WITH REPRESENTATIVE TTS SURVEYS.

Survey	Main Focus	Year
Klatt et al. [36]	Rule-based and concatenative TTS	1987
Tabet et al. [49]	Rule-based, concatenative, and parametric TTS	2011
King et al. [40]	Parametric TTS and performance measurement	2014
Tan et al. [42]	Neural, efficient, and expressive TTS	2021
Zhang et al. [50]	Diffusion-based TTS and speech enhancement	2023
Ours	Controllable TTS, datasets, metrics, and challenges	2025

B. The History of Controllable TTS

Controllable TTS aims to control various aspects of synthesized speech, such as pitch, energy, speed/duration, prosody, timbre, emotion, gender, or high-level styles. This subsection briefly reviews the history of controllable TTS, ranging from early approaches to the state-of-the-art (SOTA) in recent years.

Early Approaches. Before the prevalence of deep neural networks (DNNs), controllable TTS technologies were built primarily on rule-based, concatenative, and statistical methods. These approaches enable some degree of customization and control, however, they were constrained by the limitations of the underlying models and available computational resources. 1) Rule-based TTS systems [51]–[54], such as formant synthesis, were among the earliest methods for speech generation. These systems use manually crafted rules to simulate the speech generation process by controlling acoustic parameters such as pitch, duration, and formant frequencies, allowing explicit manipulation of prosody and phonetic details through rule adjustments. 2) Concatenative TTS [55]–[58], which dominated the field in the late 1990s and early 2000s, synthesize speech by concatenating pre-recorded speech segments, such as phonemes or diphones, stored in a large database [59]. These methods can modify the prosody by manipulating the pitch, duration, and amplitude of speech segments during concatenation. They also allow limited voice customization by selecting speech units from different speakers. 3) Parametric methods, particularly HMM-based TTS [60]–[65], gained prominence in the late 2000s. These systems model the relationships between linguistic features and acoustic parameters, providing more flexibility in controlling prosody, pitch, speaking rate, and timbre by adjusting statistical parameters. Some HMM-based systems also supported speaker adaptation [66], [67] and voice conversion [68], [69], enabling voice cloning to some extent. However, emotion can be limitedly controlled by some of these methods [60], [70]–[72]. In addition, they required less storage compared to concatenative TTS and allowed smoother transitions between speech units.

Neural Synthesis. Neural model-based TTS technologies emerged with the advent of deep learning, significantly advancing the field by enabling more flexible, natural, and expressive speech synthesis. Unlike traditional methods, neural TTS leverages DNNs to model complex relationships between input text and speech, facilitating nuanced control over various speech characteristics. Early neural TTS systems, such as WaveNet [73] and Tacotron [74], laid the groundwork for controllability. 1) Controlling prosody features like rhythm and intonation is vital for generating expressive and contextually appropriate speech. Neural TTS models achieve prosody control through explicit conditioning or learned latent representations [15], [75]–[78]. 2) Speaker control has also gained significant improvement in neural TTS through speaker embeddings or adaptation techniques [79]–[82]. 3) Besides, emotionally controllable TTS [20], [22], [31], [32], [83] has become a hot topic due to the strong modeling capability of DNNs, enabling the synthesis of speech with specific emotional tones, such as happiness, sadness, anger, or neutrality. These systems go beyond producing intelligible and natural-sounding speech, focusing on generating expressive output that aligns with the intended emotional context. 4) Neural TTS can also manipulate timbre (vocal quality) [14], [78], [84]–[87] and style (speech mannerisms) [88]–[90], allowing for creative and personalized applications. These techniques lead to one of the most popular research topics, i.e., zero-shot TTS (particularly voice cloning) [78], [82], [91], [92]. 5) Fine-grained content and linguistic control also become more powerful [93]–[96]. These methods can emphasize or de-emphasize specific words or adjust the pronunciation of phonemes through speech editing or generation techniques.

Neural TTS technologies represent a significant leap in the flexibility and quality of speech synthesis. From prosody and emotion to speaker identity and style, these systems empower diverse applications in fields such as entertainment, accessibility, and human-computer interaction.

LLM-based Synthesis. Here, we pay special attention to LLM-based synthesis methods due to their superior context modeling capabilities compared to other neural TTS methods. LLMs, such as generative pre-trained transformer (GPT) [97], [98], T5 [99], and pathways language model (PaLM) [100], have revolutionized various natural language processing (NLP) tasks with their ability to generate coherent, context-aware text. Recently, their utility has expanded into controllable TTS technologies [17], [101]–[104]. For example, users can synthesize the target speech by describing its characteristics, such as: "A young girl says 'I really like it, thank you!' with a happy voice", making speech generation significantly more intuitive and user-friendly. Specifically, an LLM can detect emotional intent in sentences (e.g., "I'm thrilled" $\rightarrow$ happiness, "This is unfortunate" $\rightarrow$ sadness). The detected emotion is encoded as an auxiliary input to the TTS model, enabling the modulation of acoustic features like prosody, pitch, and energy to align with the expressed sentiment. By leveraging LLMs' capabilities in understanding and generating rich contextual information, these systems can achieve enhanced and fine-grained control over various speech attributes, such as prosody, emotion, style, and speaker characteristics [31], [105], [106]. Integrating LLMs into TTS represents a significant step forward, enabling more dynamic and expressive speech synthesis.

C. *Organization of This Survey*

This survey first presents a comprehensive and systematic review of controllable TTS technologies, with a particular focus on model architectures, control strategies, and feature representations. To establish a foundational understanding, this survey begins with an introduction to the TTS pipeline in Section II. While our focus remains on controllable TTS, Section III examines seminal works in uncontrollable TTS that have significantly influenced the field's development. Section IV provides a thorough investigation into controllable TTS methods, analyzing both their model architectures and control strategies. Section V presents a comprehensive review of datasets and evaluation metrics. Section VI provides an in-depth analysis of the challenges encountered in achieving controllable TTS systems and identifies promising future research directions. Section VII explores the broader impacts of controllable TTS technologies, followed by the conclusion in Section VIII.

II. TTS Pipeline

In this section, we elaborate on the general pipeline that supports controllable TTS technologies, including acoustic models, speech vocoders, and feature representations. Fig. 2 depicts the general pipeline of controllable TTS, containing various model architectures and feature representations, but the control strategies will be discussed in Section IV. Readers familiar with TTS pipelines may skip ahead to Section III.

A. *Overview*

A TTS pipeline generally contains three key components, i.e., linguistic analyzer, acoustic model, and speech vocoder, where a conditional input, e.g., prompts, can be processed for controllable speech synthesis. *Linguistic analyzer* aims to extract linguistic features, e.g., phoneme duration and position, syllable stress, and utterance level, from the input text, which is a necessary step in HMM-based methods [64], [65] and a few neural model-based methods [110], [111], but is time-consuming and error-prone. *Acoustic model* is a parametric

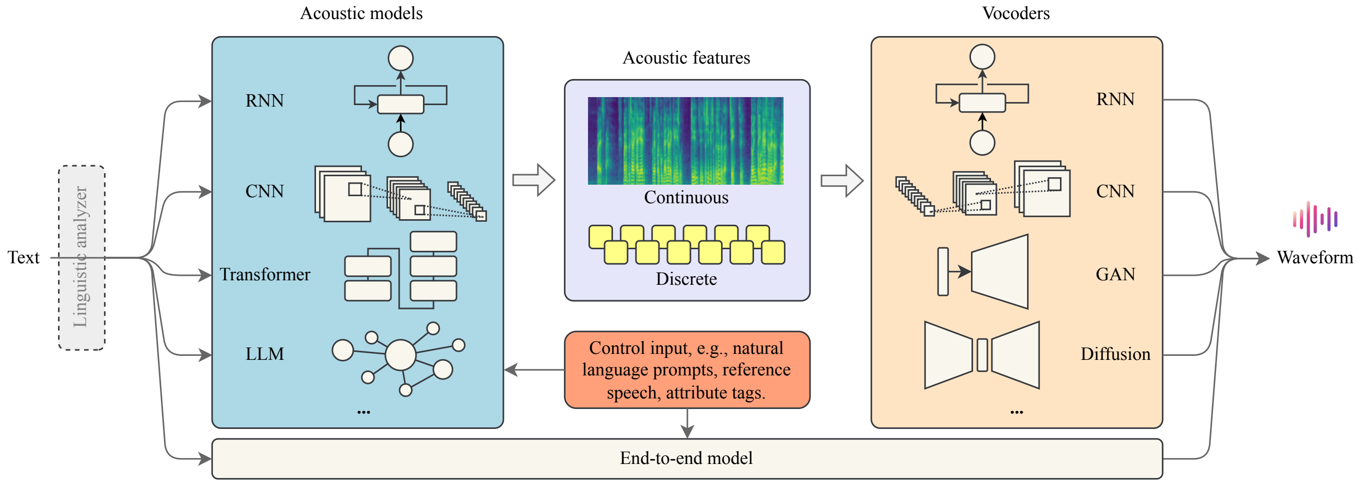


Fig. 2. General pipeline of controllable TTS from the perspective of network structure. Linguistic analysis is necessary for parametric and a few neural methods but is no longer needed for most modern neural methods. In this paper, we only review neural model-based controllable TTS methods and do not investigate acoustic features (e.g., MFCC [107], LSP [108], F0 [109]) used in early TTS methods.

or neural model that predicts the acoustic features from the input texts. Modern neural acoustic models like Tacotron [74] and later works [15], [76], [112] directly take character [113] or word embeddings [114] as the input, which is much more efficient than previous methods. *Speech vocoder* is the last component that converts the intermediate acoustic features into a waveform that can be played back. This step bridges the gap between the acoustic features and the actual sounds produced, helping to generate high-quality, natural-sounding speech [73], [115]. Besides, some end-to-end methods use a single model to encode the input and decode the speech waveforms without generating intermediate features like mel-spectrograms. Tan et al. [42] have presented a comprehensive and detailed review of acoustic models and vocoders. Therefore, the following subsections will briefly introduce some representative acoustic models and speech vocoders, followed by a discussion of acoustic feature representations.

B. Acoustic Models

Acoustic modeling is a crucial step in TTS because it ensures that the generated acoustic features capture the subtleties of human speech. By accurately modeling acoustic features, modern TTS systems can help generate high-quality and expressive audio that sounds close to human speech.

Parametric Models. Early acoustic models rely on parametric approaches, where predefined rules and mathematical functions are utilized to model speech generation. These models often utilize HMMs to capture acoustic features from linguistic input and generate acoustic features by parameterizing the vocal tract and its physiological properties such as pitch and prosody [65], [71], [72], [116]–[118]. These methods have relatively low computational costs and can produce a range of voices by adjusting model parameters. However, the speech quality of these methods is robotic and lacks natural intonation, and the expressiveness is also limited [72], [118].

RNN-based Models. Recurrent Neural Networks (RNNs) proved particularly effective in early neural TTS due to their ability to model sequential data and long-range dependencies, which helps in capturing the sequential nature of speech, such as the duration and natural flow of phonemes. Typically, these models have an encoder-decoder architecture, where an encoder encodes input linguistic features, such as phonemes or text, into a fixed-dimensional representation, and the decoder sequentially decodes this representation into acoustic features (e.g., mel-spectrogram frames) that capture the frequency and amplitude of sound over time. Tacotron 2 [75] is one of the pioneering TTS models that use RNNs with an attention mechanism, which helps align the text sequence with the generated acoustic features. It takes raw characters as input and produces mel-spectrogram frames, which are subsequently converted to waveforms. Another example is MelNet [119], which leverages autoregressive modeling to generate high-quality mel-spectrograms, demonstrating versatility in generating both speech and music, achieving high fidelity and coherence across temporal scales.

CNN-based Models. Unlike RNNs, which process sequential data frame by frame, CNNs process the entire sequence at once by applying filters across the input texts. This parallel approach enables faster training and inference, making CNN-based TTS particularly appealing for real-time and low-latency applications [16], [73], [120], [121]. Furthermore, by stacking multiple convolutional layers with varying kernel sizes or dilation rates, CNNs can capture both short-range and long-range dependencies, which are essential for natural-sounding speech synthesis. Deep Voice [16] is one of the first prominent CNN-based TTS methods, designed to generate mel-spectrograms directly from phoneme or character input. ParaNet [122] also utilizes a CNN model to achieve sequence-to-sequence mel-spectrogram generation. It uses a non-autoregressive architecture, which enables significantly faster inference by predicting multiple time steps simultaneously.

Transformer-based Models. The transformer model [123] uses self-attention layers to capture relationships within the input sequence, making them well-suited for tasks requiring an understanding of global contexts, such as prosody and rhythm in TTS. Transformer-based TTS models often employ an encoder-decoder architecture, where the encoder processes linguistic information (e.g., phonemes or text) and captures contextual relationships, and the decoder generates acoustic features (like mel-spectrograms) from these encoded representations, later converted to waveforms by a vocoder. TransformerTTS [124] is one of the first TTS models that apply transformers to synthesize speech from text. It utilizes a standard encoder-decoder transformer architecture and relies on multi-head self-attention mechanisms to model long-term dependencies, which helps maintain consistency and natural flow in speech over long utterances. FastSpeech [15] is a non-autoregressive model designed to overcome the limitations of autoregressive transformers in TTS, achieving faster synthesis than previous methods. It introduces a length regulator to align text with output frames, enabling the control of phoneme duration. FastSpeech 2 [76] extends FastSpeech by adding pitch, duration, and energy predictors, resulting in more expressive and natural-sounding speech.

LLM-based Models. LLMs [11], [26], [97], [125], known for their large-scale pre-training on text data, have shown remarkable capabilities in natural language understanding and generation. LLM-based TTS models generally use a text description to guide the mel-spectrogram generation, where the acoustic model processes the input text to generate acoustic tokens that capture linguistic and contextual information, such as tone, sentiment, and prosody. For example, PromptTTS [101] uses a textual prompt encoded by BERT [125] to guide the acoustic model on the timbre, tone, emotion, and prosody desired in the speech output. PromptTTS first generates mel-spectrograms with token embeddings and then converts them to audio using a vocoder. InstructTTS [105] generates expressive and controllable speech using natural language style prompts. It leverages discrete latent representations of speech and integrates natural language descriptions to guide the synthesis process, which bridges the gap between TTS systems and natural language interfaces, enabling fine-grained style control through intuitive prompts.

Other Acoustic Models. In TTS, generative adversarial networks (GANs) [126]–[128], variational autoencoders (VAEs) [18], [129], and diffusion models [112], [130] can also be used as acoustic models. Flow-based methods [131], [132] are also popular in acoustic feature generation. Recently proposed flow-based controllable TTS methods will be discussed in Section IV. Refer to the survey paper from Tan et al. [42] for more details.

The choice of an acoustic model depends on the specific requirements and is a trade-off between synthesis quality, computational efficiency, and flexibility. For real-time applications, CNN-based or lightweight transformer-based models are preferable, while for high-fidelity, expressive speech synthesis, transformer-based and LLM-based models are better suited.

C. Speech Vocoders

Vocoders are essential for converting acoustic features, such as mel-spectrograms, into intelligible audio waveforms and are vital in determining the naturalness and quality of synthesized speech. We broadly categorize existing vocoders according to their model architectures, i.e., RNN-, CNN-, GAN-, and diffusion-based vocoders.

RNN-based Vocoders. Unlike traditional vocoders [133], [134] that depend on manually designed signal processing pipelines, RNN-based vocoders [135]–[138] leverage the temporal modeling capabilities of RNNs to directly learn the complex patterns in speech signals, enabling the synthesis of natural-sounding waveforms with improved prosody and temporal coherence. For instance, WaveRNN [136] generates speech waveforms sample-by-sample using a single-layer recurrent neural network, typically with Gated Recurrent Units (GRU). It improves upon earlier neural vocoders like WaveNet [73] by significantly reducing the computational requirements without sacrificing audio quality. MB-WaveRNN [138] extends WaveRNN by incorporating a multi-band decomposition strategy, where the speech waveform is divided into multiple sub-bands, with each sub-band synthesized at a lower sampling rate. These sub-bands are then combined to reconstruct the full-band waveform, thereby accelerating the synthesis process while preserving audio quality.

CNN-based Vocoders. By leveraging the parallelism of convolutional operations, CNN-based vocoders [73], [139], [140] can achieve higher speech quality and more efficient synthesis compared to parametric vocoders [133], [134], making them ideal for applications that demand real-time and natural speech synthesis. However, they often require extensive training data and careful hyperparameter tuning to achieve optimal performance. WaveNet [73] is a probabilistic autoregressive model that generates waveforms sample by sample, conditioned on all preceding samples and auxiliary inputs, such as linguistic features and mel-spectrograms. It employs stacks of dilated causal convolutions, enabling long-range dependence modeling in speech signals without relying on recurrent connections. Parallel WaveNet [139] addresses WaveNet's inference speed limitations while maintaining comparable synthesis quality. It introduces a non-autoregressive mechanism based on a teacher-student framework, where the original WaveNet (teacher) distills knowledge into a student model. The student generates samples in parallel, enabling real-time synthesis without waveform quality degradation.

GAN-based Vocoders. GANs have been widely adopted in vocoders for high-quality speech generation [115], [141]–[144], leveraging adversarial losses to improve realism. GAN-based vocoders typically consist of a generator that produces waveforms conditioned on acoustic features, such as mel-spectrograms, and a discriminator that distinguishes between real and synthesized waveforms. Models like Parallel WaveGAN [143] and HiFi-GAN [115] have demonstrated the effectiveness of GANs in vocoding by introducing tailored loss functions, such as multi-scale and multi-resolution spectrogram losses, to ensure naturalness in both time and frequency domains. These models can efficiently handle the complex, non-linear relationships inherent in speech signals, resulting in high-quality synthesis. A key advantage of GAN-based vocoders is their parallel inference capability, enabling real-time synthesis with lower computational costs compared to autoregressive models. However, training GANs can be challenging due to instability and mode collapse caused by imbalanced adversarial dynamics, vanishing gradients, and the generator overfitting to limited patterns [145], [146]. Despite these challenges, GAN-based vocoders continue to advance the SOTA in neural vocoding, offering a compelling combination of speed and audio quality.

Diffusion-based Vocoders. Inspired by diffusion probabilistic models [147] that have shown success in visual generation tasks, diffusion-based vocoders [112], [148]–[151] present an alternative approach to natural-sounding speech synthesis. The core mechanism of diffusion-based vocoders involves two stages: a forward process and a reverse process. In the forward process, clean speech waveforms are progressively corrupted by adding noise in a controlled manner, creating a sequence of intermediate noisy representations. During training, the model learns to reverse this process, progressively denoising the corrupted signal to reconstruct the original waveform. Diffusion-based vocoders, such as WaveGrad [150] and DiffWave [149], have demonstrated remarkable performance in generating high-fidelity waveforms while maintaining temporal coherence and natural prosody. They offer advantages over previous vocoders, including robustness to over-smoothing [152] and the ability to model complex data distributions. However, their iterative sampling process can be computationally intensive, posing challenges for real-time applications.

Other Vocoders. There are also many other types of vocoders, such as flow-based [153]–[157] and VAE-based vocoders [122], [158], [159]. These methods provide unique strengths for speech synthesis, such as efficiency and greater flexibility in modeling complex speech variations. Readers can refer to the survey paper from Tan et al. [42] for more details.

The choice of vocoder depends on various factors. While high-quality models like GAN-based and diffusion-based vocoders excel in naturalness, they may not be suitable for real-time scenarios. On the other hand, models like Parallel WaveNet [139] balance quality and efficiency for practical use cases. The best choice will ultimately depend on the specific use case, available resources, and the importance of factors such as model size, training data, and inference speed.

D. Fully End-to-end TTS models

Fully end-to-end TTS methods [76], [159]–[162] directly generate speech waveforms from textual input, simplifying the "acoustic model → vocoder" pipeline and achieving efficient speech generation. Char2Wav [160] is an early neural TTS system that directly synthesizes speech waveforms from character-level text input. It integrates two components and jointly trains them: a recurrent sequence-to-sequence model with attention, which predicts acoustic features (e.g., mel-spectrograms) from text, and a SampleRNN-based neural vocoder [135] that generates waveforms from these features. Similarly, FastSpeech 2s [76] directly synthesizes speech waveforms from texts by extending FastSpeech 2 [76] with a waveform decoder, achieving high-quality and low-latency synthesis. VITS [159] is another fully end-to-end TTS framework. It integrates a VAE with normalizing flows [163] and adversarial training, enabling the model to learn latent representations that capture the intricate variations in speech, such as prosody and style. VITS combines non-autoregressive synthesis with stochastic latent variable modeling, achieving real-time waveform generation without compromising naturalness. There are more end-to-end TTS models such as Tacotron [74], ClariNet [161], and EATS [162], readers can refer to another survey [42] for more details. End-to-end controllable methods that emerged in recent years will be discussed in Section IV.

E. Acoustic Feature Representations

In TTS, the choice of acoustic feature representations impacts the model's flexibility, quality, expressiveness, and controllability. This subsection investigates continuous representations and discrete tokens as shown in Fig. 2, along with their pros and cons for TTS applications.

Continuous Representations. Continuous representations (e.g., mel-spectrograms and VAE features) of intermediate acoustic features use a continuous feature space to model speech signals. These representations often involve acoustic features that capture frequency, pitch, and other characteristics without discretizing the signal. The advantages of continuous features are: 1) Continuous representations retain fine-grained detail, enabling expressive and natural-sounding speech synthesis. 2) Since continuous features inherently capture variations in tone, pitch, and emphasis, they are well-suited for prosody control and emotional TTS. 3) Continuous representations are robust to information loss and can avoid quantization artifacts, allowing reconstruction of smooth audio. GAN-based [115], [143], [144] and diffusion-based methods [148], [149] often utilize continuous feature representations, i.e., mel-spectrograms. However, continuous representations are typically computationally demanding and require large-scale models and training datasets, especially in high-resolution audio synthesis.

Discrete Tokens. In discrete token-based TTS, the intermediate acoustic features (e.g., quantized units or phoneme-like tokens) are discrete values, similar to words or phonemes in languages. These are often produced using quantization techniques or learned token embeddings, such as HuBERT [166] and SoundStream [170]. The advantages of discrete tokens are: 1) Discrete tokens can encode phonemes or sub-word units, making them concise and computationally efficient to handle. 2) Discrete tokens often allow TTS systems to require fewer samples to learn and generalize, as compared with continuous representations, since the representations are compact and simplified. 3) Using discrete tokens simplifies cross-modal TTS applications like voice cloning or text prompt-based TTS, as they map well to text-like representations such as LLM tokens. LLM-based [78], [103], [105], [106] and zero-shot TTS methods [17], [78], [87] often adopt discrete tokens as their acoustic features. However, discrete representation learning may result in information loss or lack the nuanced details that can be captured in continuous representations.

Speech Quantization vs. Tokenization. It is worth noting that quantization and tokenization serve distinct purposes in speech processing. Quantization is primarily used for high-

TABLE II
POPULAR OPEN-SOURCE SPEECH QUANTIZATION AND TOKENIZATION METHODS.

Method	Modeling	Code	Year
VQ-Wav2Vec [164]	SSCP	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec#vq-wav2vec	2019
Wav2Vec 2.0 [165]	SSCP	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec	2019
HuBERT [166]	SSCP	https://github.com/facebookresearch/fairseq/tree/main/examples/hubert	2021
Whisper Encoder [167]	SSCP	https://github.com/openai/whisper	2022
Data2vec [168]	SSCP	https://github.com/facebookresearch/fairseq/tree/main/examples/data2vec	2022
W2v-BERT 2.0 [169]	SSCP	https://huggingface.co/facebook/w2v-bert-2.0	2023
SoundStream [170]	RVQ-GAN	https://github.com/wesbz/SoundStream	2021
Encodec [171]	RVQ-GAN	https://github.com/facebookresearch/encodec	2022
HiFi-Codec [172]	RVQ-GAN	https://github.com/yangdongchao/AcademiCodec	2023
SpeechTokenizer [173]	RVQ-GAN	https://github.com/ZhangXInFD/SpeechTokenizer	2023
Descript Audio Codec [174]	RVQ-GAN	https://github.com/descriptinc/descript-audio-codec	2023
Mimi Codec [175]	RVQ-GAN	https://github.com/kyutai-labs/moshi	2024
WavTokenizer [176]	VQ-GAN	https://github.com/jishengpeng/WavTokenizer	2024

SSCP: Self-supervised context (token) prediction, RVQ: Residual vector quantization [170].

fidelity compression, reducing the precision of numerical representations (e.g., from 32-bit floating point to 8-bit integers) while preserving model performance. In speech synthesis, quantization is often used in waveform generation (e.g., codec-based approaches like EnCodec [171]) and neural vocoders to compress audio signals without significant loss of perceptual quality. Tokenization, on the other hand, is a discretization process that segments continuous data into meaningful units. In speech tasks, tokenization extracts semantically relevant representations such as phonemes, characters, or learned speech units (e.g., HuBERT [166] and Wav2Vec 2.0 [165]). This makes tokenization particularly suitable for speech-to-text (ASR), TTS, and multimodal NLP tasks, where aligning speech with textual information is crucial. Tokenization also facilitates training language models on speech data by enabling linguistic or learned unit-based processing rather than raw audio waveform modeling. Table II summarizes popular open-source speech quantization and tokenization methods. Tables IV and III summarize the types of acoustic features of representative methods.

III. UNCONTROLLABLE TTS

The development of uncontrollable text-to-speech (UC-TTS) systems represents a significant shift from traditional, linguistics-based synthesis to modern, data-driven deep learning techniques. This shift highlights the integration of both local and global information to produce speech with human-like quality and naturalness. This section briefly investigates UC-TTS methods, emphasizing the role of local and global information in enhancing speech fidelity and expressiveness.

In the context of UC-TTS, the term “uncontrollable” refers to the absence of explicit control mechanisms for speech features such as emotion, timbre, and speaking style. Despite this, the goal is to achieve natural, fluid speech while mitigating issues like mispronunciations and omissions.

A. Early Approaches: Statistical Models

Early TTS systems relied on statistical models such as HMMs [64], [65] and early neural network-based parametric methods [110], [111]. These models operated at the frame level, using acoustic models and vocoders for text-to-speech conversion. Notable contributions from Tokuda et al. [177] employed HMMs for statistical parametric synthesis, focusing on local features like phonemes, accents, and prosody to improve speech naturalness.

While robust, these statistical methods were limited by their reliance on pre-segmented data, leading to oversimplified assumptions about speech dynamics. Local linguistic features were well-modeled, but the global phonetic context was often overlooked, resulting in speech that sounded monotone and lacked emotional depth, as noted by Zen et al. [41].

B. Sequence-to-Sequence Models

The emergence of sequence-to-sequence models represents a significant breakthrough by removing the need for explicit linguistic features, thereby enabling the capture of the nuances of human speech. Models such as Tacotron [74] and Tacotron 2 [75] utilize RNNs with attention mechanisms to effectively model the complex, nonlinear nature of speech sequences. These innovations enable the tuning of speech parameters, enhancing prosody and rhythm by modeling entire utterances instead of isolated phonetic units.

Building on these advancements, Deep Voice 3 [121] introduces a fully convolutional sequence-to-sequence architecture that significantly accelerates training speed compared to RNN-based models. This approach achieves training times an order of magnitude faster, enabling scalability to handle large datasets. Additionally, the position-augmented attention mechanism in Deep Voice 3 enhances the naturalness of synthesized speech, achieving competitive mean opinion scores, especially when paired with advanced neural vocoders like WaveNet. This development not only improves training efficiency but also enhances the scalability and naturalness of TTS systems.

C. Transformer-based Models

Transformer-based architectures advanced the field by enabling computational parallelization and effectively capturing long-range dependencies. Models like Transformer TTS overcame RNN challenges, such as gradient vanishing, by using efficient training paradigms [124]. Self-attention mechanisms allowed simultaneous modeling of local phonetic details and global prosodic contexts, resulting in more sophisticated and human-like speech synthesis.

Although transformers improved contextual information incorporation, challenges remained in preserving local phonetic precision. To address these, techniques such as relative position encodings and localized attention were integrated [123].

D. Integrating Flow and Diffusion Models

Recent advancements have shifted toward integrating global information within end-to-end architectures to enhance speech naturalness and coherence. Flow-based models like Glow-TTS [132] and Flow-TTS [131] exemplify this by employing invertible transformations that maintain the balance between local precision and global coherence. These architectures enable the synthesis of high-fidelity speech by modeling complex dependencies across the entire utterance, thus improving the overall fluidity and naturalness of the generated speech.

Moreover, the introduction of diffusion models in TTS, such as WaveGrad 2 [178], highlights the shift toward models that can iteratively refine speech output. These models use score-matching and diffusion processes to generate speech directly from phoneme sequences, effectively capturing both local nuances and overarching global patterns. The iterative nature of these models allows for adjustments that enhance the quality of the synthesized audio, accommodating variations in speech without explicit control over specific attributes.

The integration of adversarial training and VAEs further exemplifies the evolution toward incorporating global information. Systems like VITS [159] leverage these techniques to enhance expressiveness and naturalness by learning complex mappings between text and speech. This approach allows the model to manage variations in prosody and rhythm that are inherently derived from the textual input, aligning with the objectives of UC-TTS to produce diverse and natural speech.

The evolution from HMMs to advanced architectures in UC-TTS exemplifies progress toward synthesizing speech that is both expressive and precise. The interplay of local and global information is crucial for enhancing speech quality and customizability. Future UC-TTS research aims to produce high-fidelity, customizable speech by harmonizing deep contextual insights with precise local adjustments, meeting diverse user needs and communication contexts.

IV. Controllable TTS

In this section, we first review recent TTS work from the perspective of model architecture, followed by a detailed discussion of control strategies in controllable TTS, which is the core part of this survey. Current model architectures can be broadly classified into two main categories: The first is the non-autoregressive (NAR) generative models, which are based on HMMs, neural networks, VAEs, diffusion models, flow matching, and other NAR techniques. The second category relies on autoregressive (AR) codec language models, which typically quantize speech into discrete tokens and use decoder-only models to autoregressively generate these tokens. We summarize the NAR-based and AR-based controllable TTS methods in Table III and Table IV, respectively.

A. Non-Autoregressive Architectures

In non-autoregressive TTS models, the model generates the entire output sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ at once, conditioned on the input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$. The probability distribution for generating the sequence is:

$$P(\mathbf{y}|\mathbf{x}) = P(\mathbf{y}|\mathbf{x}, \theta), \tag{1}$$

Where $P(\mathbf{y}|\mathbf{x})$ is the likelihood of the output sequence $\mathbf{y}$ given the input $\mathbf{x}$, and θ represents the parameters of the model (e.g., weights). Since the output sequence is predicted simultaneously, the model learns to capture dependencies in a way that does not rely on previously generated outputs.

Transformer-based Approaches. Advancements in controllable TTS technology highlight the integration of deep learning with audio processing, driven by Transformer-based architectures. Ren et al. [15] introduced FastSpeech, a feed-forward non-autoregressive Transformer model that significantly enhances TTS efficiency by reducing inference time and improving the stability issues found in autoregressive models like Tacotron 2. This model provides precise control over prosodic features through duration prediction, effectively tackling the one-to-many mapping challenge. FastSpeech 2 [180] builds on this by integrating pitch and energy control, eliminating the need for the complex teacher-student distillation process, thus enhancing training efficiency and improving voice quality. Parallel Tacotron [84] further advances TTS by employing a variational autoencoder-based residual encoder, capturing intricate prosodic nuances. This approach, combined with iterative spectrogram loss, significantly enhances the naturalness and quality of synthesized speech. Additionally, FastPitch [77] incorporates direct pitch prediction into its architecture, enabling fully parallelized synthesis and precise pitch manipulation. This capability enhances expressiveness and retains the efficiency benefits established by FastSpeech. These innovations significantly contribute to the development of more interactive and natural AI-driven communication systems, underscoring the potential of integrating AI with human-centric disciplines to craft a future where technology and humanity coexist harmoniously.

VAE-based Approaches. Recent advancements in controllable TTS systems are largely driven by the integration of VAE architectures, which enhance the flexibility and precision of speech modulation. Zhang et al. [18] pioneered the use of VAEs in end-to-end speech synthesis, creating disentangled latent representations that allow effective style control and transfer, especially in prosody and emotion management, outperforming the Global Style Token (GST) model [19] in style transfer tasks. Building on this, Hsu et al. [129] developed a hierarchical generative model with a conditional VAE framework and a Gaussian mixture model, enabling precise control over complex speech attributes such as environment and style, thus improving expressive speech synthesis through refined noise and speaker characteristic management. Liu et al. [188] further advanced the field with the CLONE model, a single-stage TTS system that resolves the one-to-many mapping issue and enhances high-frequency information reconstruction. By employing a conditional VAE with normalizing flows

TABLE III
A SUMMARY OF EXISTING NON-AUTOREGRESSIVE CONTROLLABLE NEURAL-BASED METHODS.

Method	Zero-shot TTS	Controlability								Model Architectures		Acoustic Feature	Release Time
		Pit.	Ene.	Spe.	Pro.	Tim.	Emo.	Env.	Des.	Acoustic Model	Vocoder		
FastSpeech [15]				✓	✓					Transformer	WaveGlow [179]	MelS	2019.05
FastSpeech 2 [180]		✓	✓	✓	✓					Transformer	Parallel WaveGAN [143]	MelS	2020.06
FastPitch [77]		✓			✓					Transformer	WaveGlow	MelS	2020.06
Parallel Tacotron [181]					✓					Transformer + CNN	WaveRNN [136]	MelS	2020.10
StyleTagging-TTS [182]	✓					✓	✓			Transformer + CNN	HiFi-GAN [115]	MelS	2021.04
SC-GlowTTS [183]	✓					✓				Transformer + Flow	HiFi-GAN	MelS	2021.06
Meta-StyleSpeech [184]	✓					✓				Transformer	MelGAN [144]	MelS	2021.06
DelightfulTTS [185]		✓		✓	✓					Transformer + CNN	HiFiNet [185]	MelS	2021.11
YourTTS [82]	✓					✓				Transformer + Flow	HiFi-GAN	LinS	2021.12
StyleTTS [88]	✓					✓				CNN + RNN	HiFi-GAN	MelS	2022.05
GenerSpeech [90]	✓					✓				Transformer + Flow	HiFi-GAN	MelS	2022.05
Cauliflow [186]				✓	✓					BERT + Flow	UP WaveNet [187]	MelS	2022.06
CLONE [188]		✓		✓	✓					Transformer + CNN	WaveNet [73]	MelS + LinS	2022.07
PromptTTS [101]		✓	✓	✓		✓	✓		✓	BERT + Transformer	HiFi-GAN	MelS	2022.11
Grad-StyleSpeech [189]	✓					✓				Score-based Diffusion	HiFi-GAN	MelS	2022.11
NaturalSpeech 2 [86]	✓					✓				Diffusion	RVQ-based Codec [86]	Latent Feature	2023.04
PromptStyle [190]	✓	✓			✓	✓	✓		✓	VITS + Flow	HiFi-GAN	MelS	2023.05
StyleTTS 2 [89]	✓				✓	✓	✓			Flow-based Diffusion + GAN	HifiGAN / iSTFTNet [191]	MelS	2023.06
VoiceBox [192]	✓					✓				Transformer + Flow	HiFi-GAN	MelS	2023.06
MegaTTS 2 [193]	✓				✓	✓	✓			Decoder-only Transformer + GAN	HiFi-GAN	MelS	2023.07
PromptTTS 2 [102]		✓	✓	✓		✓			✓	Diffusion	RVQ-based Codec	Latent Feature	2023.09
VoiceLDM [194]		✓			✓	✓	✓	✓	✓	Diffusion	HiFi-GAN	MelS	2023.09
DurIAN-E [195]		✓		✓	✓					CNN + RNN	HiFi-GAN	MelS	2023.09
PromptTTS++ [104]		✓		✓	✓	✓	✓		✓	Transformer + Diffusion	BigVGAN [196]	MelS	2023.09
SpeechFlow [197]	✓					✓				Transformer + Flow	HiFi-GAN	MelS	2023.10
P-Flow [198]	✓					✓				Transformer + Flow	HiFi-GAN	MelS	2023.10
E3 TTS [199]	✓					✓				Diffusion	Not required	Waveform	2023.11
HierSpeech++ [200]	✓					✓				Transformer + VAE + Flow	BigVGAN	MelS	2023.11
Audiobox [201]	✓	✓		✓	✓	✓		✓	✓	Transformer + Flow	EnCodec [171]	MelS	2023.12
FlashSpeech [202]	✓					✓				Latent Consistency Model	EnCodec	Token	2024.04
NaturalSpeech 3 [87]	✓			✓	✓	✓				Transformer + Diffusion	FACodec [87]	Token	2024.04
InstructTTS [105]		✓		✓	✓	✓	✓		✓	Transformer + Diffusion	HiFi-GAN	Token	2024.05
ControlSpeech [106]	✓	✓	✓	✓	✓	✓	✓		✓	Transformer + Diffusion	FACodec	Token	2024.06
AST-LDM [203]						✓		✓	✓	Diffusion + VAE	HiFi-GAN	MelS	2024.06
SimpleSpeech [204]	✓					✓				Transformer + Diffusion	SQ Codec [204]	Token	2024.06
DiTTo-TTS [205]	✓			✓		✓				DiT + VAE	BigVGAN	MelS	2024.06
E2 TTS [206]	✓					✓				Transformer + Flow	BigVGAN	MelS	2024.06
MobileSpeech [207]	✓					✓				Transformer	Vocos [208]	Token	2024.06
DEX-TTS [209]	✓					✓				Diffusion	HiFi-GAN	MelS	2024.06
ArtSpeech [210]	✓					✓				RNN + CNN	HiFI-GAN	MelS + Energy + F0	2024.07
CCSP [211]	✓					✓				Diffusion	RVQ-based Codec [211]	Token	2024.07
SimpleSpeech 2 [212]	✓			✓		✓				Flow-based DiT	SQ Codec	Token	2024.08
E1 TTS [213]	✓					✓				DiT + Flow	BigVGAN	Token + MelS	2024.09
StyleTTS-ZS [214]	✓					✓				Flow-based Diffusion + GAN	Mel-based Decoder [214]	MelS	2024.09
NansyTTS [215]	✓	✓		✓	✓	✓			✓	Transformer	NANSY++ [215]	MelS	2024.09
NanoVoice [216]	✓					✓				Diffusion	BigVGAN	MelS	2024.09
MS^{2}KU-VTTS [217]								✓	✓	Transformer	BigVGAN	MelS	2024.10
MaskGCT [78]	✓			✓		✓				Transformer + Flow	Vocos	Token	2024.10
EmoSphere++ [218]	✓				✓	✓	✓			Transformer + Flow	BigVGAN	MelS	2024.11
EmoDubber [219]	✓				✓	✓	✓			Transformer + Flow	Flow-based Vocoder [219]	MelS	2024.12
HED [220]	✓						✓			Flow-based Diffusion	Vocos	MelS	2024.12
DiffStyleTTS [221]		✓	✓	✓	✓	✓				Transformer + Diffusion	HiFi-GAN	MelS	2025.01
DrawSpeech [222]			✓		✓					Diffusion	HiFi-GAN	MelS	2025.01
ProEmo [223]		✓	✓				✓		✓	Transformer	HiFi-GAN	MelS	2025.01

Abbreviations: Pit(ch), Ene(rgy), Spe(ed), Pro(sody), Tim(bre), Emo(tion), Env(ironment), Des(cription). MelS: Mel Spectrogram. LinS: Linear Spectrogram.

and a dual-path adversarial training mechanism with multi-band discriminators, CLONE achieves nuanced control over prosody and energy, demonstrating superior performance in both speech quality and prosody control compared to state-of-the-art models. These collective innovations highlight the adaptability of VAEs in managing complex speech generation tasks, marking significant progress toward more dynamic and versatile TTS technologies, with ongoing research promising even greater advancements.

Diffusion-based Approaches. The core concept of diffusion-based models is to generate target data by progressively removing noise. During the forward diffusion phase, noise is incrementally added to the original data to form a noise distribution. In the generation phase, a reverse denoising process is employed to gradually recover high-quality speech from the noise. Grad-StyleSpeech [189] introduces a hierarchical transformer encoder to create a representative noise prior distribution for speaker-adaptive settings using score-based diffusion models. NaturalSpeech 2 [86] uses a neural audio codec with residual vector quantizers to obtain quantized latent vectors, which are then generated using a diffusion model conditioned on text input. NaturalSpeech 3 [87] decomposes speech into distinct subspaces that represent different attributes and generates each subspace independently. DEX-TTS [209] improves DiT-based diffusion networks by applying overlapping patching and convolution-frequency patch embedding strategies. E3 TTS [199] models the temporal structure of the waveform through the diffusion process, eliminating the need for any intermediate representations, such as spectrogram features or alignment information.

Applying diffusion models to TTS requires a complex pipeline due to the need for precise temporal alignment between text and speech and the high fidelity required for audio data. This includes domain-specific modeling, such as phoneme and duration [86]. To address the issue of reduced naturalness caused by the addition of duration models, DiTTo-TTS [205] leverages the off-the-shelf pre-trained text and speech encoders without relying on speech domain-specific modeling by incorporating cross-attention mechanisms with the prediction of the total length of speech representations. Similarly, SimpleSpeech [204] proposes a speech codec model (SQ-Codec) based on scalar quantization and uses the sentence duration to control the generated speech length. Besides TTS, some text-to-audio models can also perform TTS and are worth referring to, such as AudioLDM [224], AudioLDM2 [225], Make-An-Audio [226], and CosyAudio [227].

Flow-based Approaches. Flow-based methods leverage invertible flow transformations [163], [228] to learn mappings from target speech features to simple distributions [179], typically standard Gaussian distributions. Due to their invertibility, this mechanism can directly sample from the simple distribution and generate high-fidelity speech in the reverse direction. Audiobox [201] and P-flow [198] employ non-autoregressive flow-matching models [228] for efficient and stable speech synthesis. VoiceBox [192] also employs flow-matching to generate speech, effectively casting the TTS task into a speech infilling task. SpeechFlow [197] is trained on 60k hours of untranscribed speech with flow matching and mask conditions and can be fine-tuned with task-specific data to match or surpass existing expert models. This highlights the potential of generative models as foundation models for speech applications. HierSpeech++ [200] proposes a hierarchical variational inference method. FlashSpeech [202] is built on a latent consistency model and applies a novel adversarial consistency training approach that can train from scratch without the need for a pre-trained diffusion model as the teacher, achieving speech generation in one or two steps.

Recently, E2 TTS [206] converts text input into a character sequence with filler tokens and trains a mel spectrogram generator based on an audio infilling task, achieving human-level naturalness. Inspired by E2 TTS, F5-TTS [229] refines the text representation with ConvNext v2 [230], facilitating easier alignment with speech. E1 TTS [213] further distills a diffusion-based TTS model into a one-step generator with distribution matching distillation [231], [232], reducing the number of network evaluations in sampling from diffusion models. SimpleSpeech 2 [212] introduces a flow-based scalar transformer diffusion model. The work also provides a theoretical analysis, showing that the inclusion of a small number of noisy labels in a large-scale dataset is equivalent to introducing classifier-free guidance during model optimization.

Other NAR Approaches. Other works leverage GAN-based or masked generative model-based methods for TTS generation. StyleTTS 2 [89] employs large pre-trained speech language models (SLMs) such as Wav2Vec 2.0 [165], HuBERT [166], and WavLM [233] as discriminators in combination with a novel differentiable duration modeling approach. This setup uses SLM representations to enhance the naturalness of the synthesized speech. MaskGCT [78] proposes masked generative transformers without requiring text-speech alignment supervision and phone-level duration prediction. The model employs a two-stage system trained using a mask-and-predict learning paradigm.

B. Autoregressive Architectures

For an autoregressive model in TTS, the probability of the speech frame sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ given the input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ can be modeled as:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_{<t}, \mathbf{x}), \tag{2}$$

where y_t is the predicted output frame at time step t, $y_{<t} = (y_1, \ldots, y_{t-1})$ are the previous frames, and $\mathbf{x}$ is the input feature sequence. Each frame y_t is predicted conditioned on all previous frames and the input sequence $\mathbf{x}$. Autoregressive models are powerful in TTS modeling but tend to be slower in generation time compared to non-autoregressive models, making them suitable for applications where quality is prioritized over real-time performance.

HMM-based Approaches. In the realm of controllable TTS, advancements in HMM architectures have significantly enhanced the manipulation of speech elements such as emotion and prosody. Yamagishi et al. [70] pioneered this field by introducing style-dependent and style-mixed modeling, which allowed precise emulation of human-like emotional nuances and versatile synthesis across various styles by incorporating style as a contextual variable. Building on this foundation, Qin et al. [268] developed the "average emotion model," which utilized maximum likelihood linear regression-based adaptation to modulate emotions like happiness and sadness even with limited data, thus advancing the emotional intelligence of synthetic speech systems.

Furthering expressive variability, Nose et al. [117] integrated subjective style intensities and a multiple-regression global variance model into HMMs, addressing over-smoothing and enabling nuanced emotional expressions. Lorenzo-Trueba et al. [72] expanded on these capabilities with CSMAPLR [72] adaptation, introducing "emotion transplantation" to transfer emotional states between speakers while preserving voice distinctiveness, enhancing personalized human-computer interaction. These innovations in HMM architectures have broadened the expressiveness and individuality in synthetic speech, augmenting technological interfaces and paving the way for future developments in adaptive, lifelike speech generation.

RNN-based Approaches. Controllable TTS technology has seen significant advancements through innovations in neural network architectures, particularly RNN-based architectures, enabling natural-sounding speech generation with adjustable emotion, prosody, and pitch. Prosody-Tacotron [234] is an extension of the original Tacotron model, designed to improve the prosody (rhythm, intonation, stress patterns) of synthesized speech. It builds on the Tacotron framework by introducing additional mechanisms to explicitly control prosodic features, which is a key challenge in TTS systems. Wang et al. [19] introduced global style tokens, using an unsupervised approach to encapsulate diverse speech styles into fixed tokens, thus enabling versatile style transfer within the Tacotron framework.

TABLE IV
A SUMMARY OF EXISTING AUTOREGRESSIVE CONTROLLABLE NEURAL-BASED METHODS.

Method	Zero-shot TTS	Controlability								Model Architectures		Acoustic Feature	Release Time
		Pit.	Ene.	Spe.	Pro.	Tim.	Emo.	Env.	Des.	Acoustic Model	Vocoder		
Prosody-Tacotron [234]		✓			✓					RNN	WaveNet	MelS	2018.03
GST-Tacotron [235]		✓			✓					CNN + RNN	Griffin-Lim	LinS	2018.03
GMVAE-Tacotron [129]		✓		✓	✓			✓		VAE	WaveRNN	MelS	2018.12
VAE-Tacotron [18]		✓		✓	✓					CNN + RNN	WaveNet	MelS	2019.02
DurIAN [138]		✓		✓	✓					CNN + RNN	Multi-band WaveRNN [138]	MelS	2019.09
Flowtron [236]		✓		✓	✓					CNN + RNN	WaveGlow	MelS	2020.07
MsEmoTTS [83]		✓			✓		✓			CNN + RNN	WaveRNN	MelS	2022.01
VALL-E [85]	✓					✓				Decoder-only Transformer	EnCodec	Token	2023.01
SpearTTS [237]	✓					✓				Decoder-only Transformer	SoundStream [170]	Token	2023.02
VALL-E X [238]	✓					✓				Decoder-only Transformer	EnCodec	Token	2023.03
Make-A-Voice [239]	✓					✓				Encoder-decoder Transformer	Unit-based Vocoder [239]	Token	2023.05
TorToise [240]						✓				Decoder-only Transformer + Diffusion	UnivNet [241]	MelS	2023.05
MegaTTS [91]	✓					✓				Decoder-only Transformer + GAN	HiFi-GAN	MelS	2023.06
SC VALL-E [242]	✓	✓		✓	✓	✓	✓			Decoder-only Transformer	EnCodec	Token	2023.07
Salle [243]		✓	✓	✓	✓	✓	✓		✓	Decoder-only Transformer	EnCodec	Token	2023.08
UniAudio [244]	✓	✓		✓	✓	✓			✓	Decoder-only Transformer	EnCodec	Token	2023.10
ELLA-V [245]	✓					✓				Decoder-only Transformer	EnCodec	Token	2024.01
Base TTS [246]	✓					✓				Decoder-only Transformer	HiFi-GAN+BigVGAN	Token	2024.02
CLaM-TTS [247]	✓					✓				Encoder-decoder Transformer	BigVGAN	Token + MelS	2024.04
RALL-E [248]	✓					✓				Decoder-only Transformer	SoundStream	Token	2024.05
ARDiT [249]	✓			✓		✓				Decoder-only DiT	BigVGAN	MelS	2024.06
VALL-E R [250]	✓					✓				Decoder-only Transformer	Vocos	Token	2024.06
VALL-E 2 [251]	✓					✓				Decoder-only Transformer	Vocos	Token	2024.06
Seed-TTS [252]	✓					✓	✓			Decoder-only Transformer + DiT	Unknown	Latent Feature	2024.06
VoiceCraft [93]	✓					✓				Decoder-only Transformer	HiFi-GAN	Token	2024.06
XTTS [253]	✓					✓				Decoder-only Transformer	HiFi-GAN-based Vocoder [253]	Token + MelS	2024.06
CosyVoice [17]	✓	✓		✓	✓	✓	✓		✓	Decoder-only Transformer + Flow	HiFi-GAN	Token	2024.07
MELLE [254]	✓					✓				Decoder-only Transformer	HiFi-GAN	MelS	2024.07
VoxInstruct [103]	✓	✓	✓	✓	✓	✓	✓		✓	Decoder-only Transformer	Vocos	Token	2024.08
Emo-DPO [31]							✓			Decoder-only Transformer	HiFi-GAN	Token + MelS	2024.09
FireRedTTS [255]	✓				✓	✓				Decoder-only Transformer + Flow	BigVGAN	Token + MelS	2024.09
CoFi-Speech [256]	✓					✓				Decoder-only Transformer	BigVGAN	Token + MelS	2024.09
Takin [257]	✓	✓		✓	✓	✓	✓		✓	Decoder-only Transformer + Flow	HiFi-GAN	Token + MelS	2024.09
HALL-E [258]	✓					✓				Decoder-only Transformer	EnCodec	Token	2024.10
FishSpeech [259]	✓					✓				Decoder-only Transformer	Firefly-GAN [259]	Token	2024.11
SLAM-Omni [260]	✓				✓	✓				Decoder-only Transformer	HiFi-GAN	Token + MelS	2024.12
IST-LM [261]	✓				✓	✓				Decoder-only Transformer	HiFi-GAN	Token + MelS	2024.12
KALL-E [262]	✓				✓	✓	✓			Decoder-only Transformer	WaveVAE [262]	Latent Feature	2024.12
IDEA-TTS [263]	✓					✓		✓		Transformer	Flow-based Vocoder [263]	LinS + MelS	2024.12
FleSpeech [264]	✓	✓	✓	✓	✓	✓	✓		✓	Flow-based DiT	WaveGAN [141]	Latent Feature	2025.01
Step-Audio [265]	✓				✓	✓	✓		✓	Decoder-only Transformer	Flow-based Vocoder [265]	Token	2025.02
Vevo [266]	✓				✓	✓	✓			Decoder-only Transformer	BigVGAN	Token + MelS	2025.02
Spark-TTS [267]	✓	✓		✓	✓	✓				Decoder-only Transformer	BiCodec [267]	Token	2025.03

Abbreviations: Pit(ch), Ene(rgy), Spe(ed), Pro(sody), Tim(bre), Emo(tion), Env(ironment), Des(cription). MelS: Mel Spectrogram. LinS: Linear Spectrogram.

Similarly, Stanton et al. [235] introduce Text-Predicted Global Style Tokens (TP-GST), enhancing Tacotron speech synthesis to predict expressive speaking styles directly from text without auxiliary inputs. Experiments show TP-GST generates more prosodic variation and achieves higher listener preference than baselines. Skerry-Ryan et al. [234] further advanced this by incorporating prosodic embeddings, providing control over timing and intonation, and significantly improving the replication of emotions in synthetic speech.

Building on these innovations, emotion-controllable models developed by Li et al. [21] focus on calibrating emotional nuances using emotion embedding networks and style loss alignment, allowing detailed modulation of emotional strength. Hierarchical models like MsEmoTTS [83] refine this approach by segmenting synthesis into global, utterance-level, and local emotional strengths, offering enhanced emotional expressiveness and intuitive control. These advancements have expanded the scope to produce nuanced TTS outputs, enabling precise control over emotion, prosody, and pitch, with applications ranging from virtual assistants to interactive narratives. As researchers continue to explore the potential of neural networks in TTS, the technology promises even richer, more engaging digital experiences, moving towards speech synthesis that is indistinguishable from natural human interaction.

LLM-based Approaches. Inspired by the success of LLMs in natural language processing (NLP), recent studies have explored leveraging in-context learning for zero-shot TTS. As shown in Fig. 3, LLM-based approaches often take the target text or instructions and an optional reference speech clip as input, and utilize autoregressive modeling to generate speech tokens or features, which are then converted into the final waveform by a decoder.

VALL-E [85] is a pioneering work in this area, formulating TTS as a conditional language modeling problem. It utilizes EnCodec [269] to discretize waveforms into tokens as intermediate representations and employs a two-stage modeling pipeline: an autoregressive model first generates coarse audio tokens, followed by a non-autoregressive model that iteratively predicts additional codebook codes for refinement. This hierarchical modeling of semantic and acoustic tokens has set the foundation for many subsequent LLM-based TTS approaches [237], [239], [242], [257].

Building on VALL-E, various improvements have been proposed. VALL-E X [238] extends VALL-E to multilingual scenarios, supporting zero-shot cross-lingual speech synthesis and speech-to-speech translation. ELLA-V [245] intro-

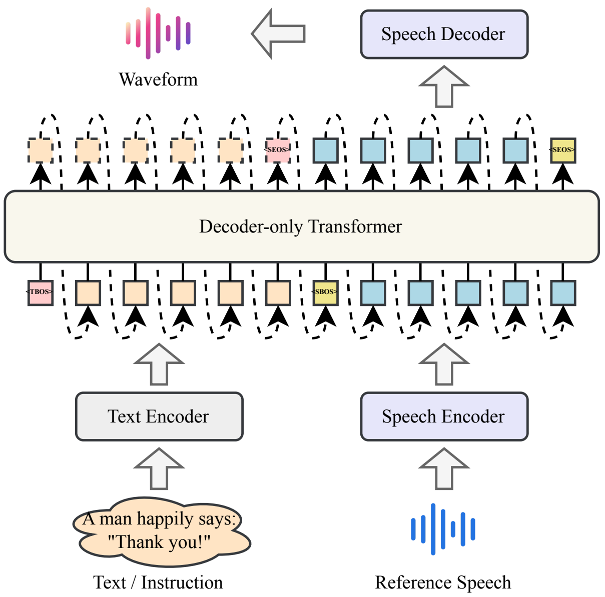


Fig. 3. The typical architecture of LLM-based TTS methods. Orange and blue squares are text and speech tokens, respectively. Pink and yellow squares are $<$ BOS $>$ and $<$ EOS $>$ tokens for text and speech sequences.

duces a sequence order rearrangement step, enhancing local alignment between phoneme and acoustic modalities. RALL-E [248] incorporates prosody tokens as chain-of-thought prompting [270] to stabilize the generation of speech tokens. VALL-E R [250] improves phoneme-to-acoustic alignment and adopts codec-merging to boost decoding efficiency and reduce computational overhead. VALL-E 2 [251] introduces repetition-aware sampling and grouped code modeling for greater stability and faster inference. HALL-E [258] adopts a hierarchical post-training framework, effectively managing the trade-off between reducing the frame rate and producing high-quality speech.

Beyond the foundational improvements introduced by VALL-E and its extensions, further advancements have focused on enhancing speech alignment, quality, and robustness. SpearTTS [237] and Make-a-Voice [239] use semantic tokens to bridge the gap between text and acoustic features. FireRedTTS [255] further optimizes the tokenizer architecture to enhance speech quality. CoFi-Speech [256] generates speech in a coarse-to-fine manner via a multi-scale speech coding and generation approach, producing natural and intelligible speech. Similarly, BASE TTS [246] introduces discrete speech representations based on the WavLM [233] self-supervised model, focusing on phonemic and prosodic information. SeedTTS [252] also proposes a self-distillation method for speech decomposition and a reinforcement learning approach to enhance the robustness, speaker similarity, and controllability of generated speech.

Although models using discrete tokens as intermediate representations have achieved notable success in zero-shot TTS, they still face fidelity issues compared to continuous representations like Mel spectrograms [249], [254]. MELLE [254] optimizes the training objectives and sampling strategy, marking the first exploration of using continuous-valued tokens instead of discrete-valued tokens within the paradigm of autoregressive speech synthesis models. Similar to MELLE, ARDiT [249] encodes audio as a vector sequence in continuous space and autoregressively generates these sequences by a decoder-only transformer.

TABLE V
A SUMMARY OF CONTROL STRATEGIES FOR TTS.

Control Strategy	Control Signal
Style Tagging	Discrete labels, continuous values, latent variables, speech audio
Speech Reference Prompt	Speech audio
Natural Language Descriptions	User input text, speech audio
Instruction-Guided Control	User input text, speech audio

Additionally, some autoregressive methods enable speech editing through natural language instructions. VoiceCraft [93] introduces a decoder-only Transformer-based neural codec language model that utilizes causal masking and delayed stacking for token rearrangement. This approach allows for bidirectional context-aware speech editing and zero-shot TTS. VoiceCraft achieves precise control through text-guided modifications, such as insertion, deletion, and substitution, producing edits that are nearly indistinguishable from unmodified recordings in terms of naturalness. InstructSpeech [271] employs a multi-task LLM trained on triplet data (instruction, input, and output speech) with task embeddings and hierarchical adapters. It enables fine-grained control over both semantic attributes (content editing) and acoustic properties (emotion, speed) using natural language instructions. By leveraging multi-step reasoning, InstructSpeech facilitates free-form editing and efficiently adapts to new tasks.

C. Control Strategies

The control strategies in existing controllable TTS can be broadly classified into four categories: style tagging using discrete labels or continuous control signals, reference speech prompt for customizing a new speaker's voice with just a few seconds of voice input, controlling speech style using natural language descriptions, and instruction-guided speech attributes control. We illustrate taxonomies of controllable TTS from the perspective of control strategies in Fig. 4.

Style Tagging. This paradigm enables controllable speech synthesis by adjusting key attributes such as pitch, energy, speech rate, and emotion. These attributes can be controlled using either categorical labels or continuous values. In this context, "tagging" refers to the process of assigning a control signal to a specific speech attribute. Some approaches utilize discrete labels to control synthesized speech. For instance, StyleTagging-TTS [182] employs short phrases or words to represent utterance styles. It learns the relationship between linguistic embeddings and style embeddings using a pre-trained language model. Emo-DPO [31] enables precise emotion control (e.g., "angry," "happy") via Direct Preference Optimization (DPO) [276] with LLM. By training on paired emotion-text data, it leverages contrastive learning to distinguish subtle prosodic differences between emotions. Users can specify an emotion label to shape the expressive quality of

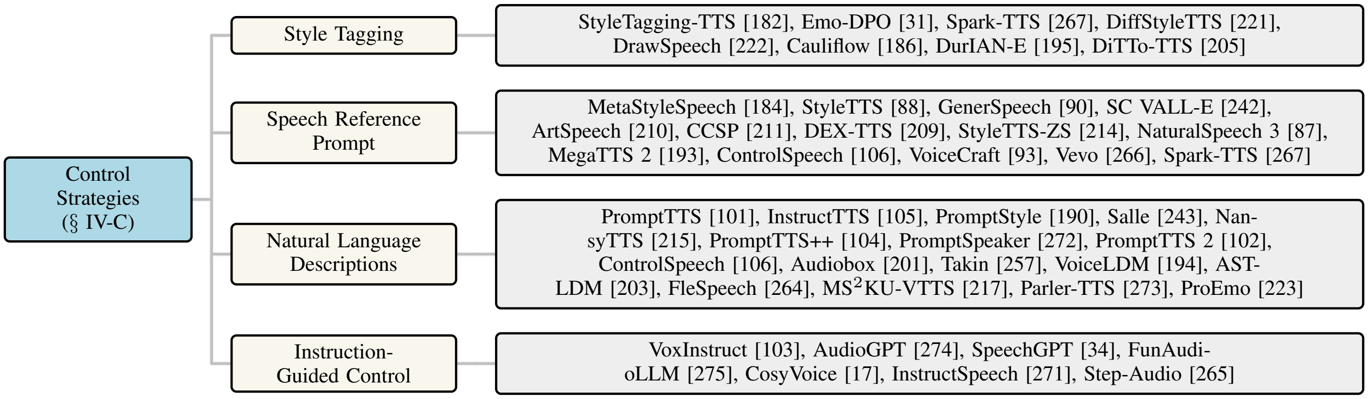

Control Strategies (§ IV-C)
Style Tagging
StyleTagging-TTS [182], Emo-DPO [31], Spark-TTS [267], DiffStyleTTS [221], DrawSpeech [222], Cauliflow [186], DurIAN-E [195], DiTTo-TTS [205]
Speech Reference Prompt
MetaStyleSpeech [184], StyleTTS [88], GenerSpeech [90], SC VALL-E [242], ArtSpeech [210], CCSP [211], DEX-TTS [209], StyleTTS-ZS [214], NaturalSpeech 3 [87], MegaTTS 2 [193], ControlSpeech [106], VoiceCraft [93], Vevo [266], Spark-TTS [267]
Natural Language Descriptions
PromptTTS [101], InstructTTS [105], PromptStyle [190], Salle [243], NansyTTS [215], PromptTTS++ [104], PromptSpeaker [272], PromptTTS 2 [102], ControlSpeech [106], Audiobox [201], Takin [257], VoiceLDM [194], AST-LDM [203], FleSpeech [264], MS^{2}KU-VTTS [217], Parler-TTS [273], ProEmo [223]
Instruction-Guided Control
VoxInstruct [103], AudioGPT [274], SpeechGPT [34], FunAudioLLM [275], CosyVoice [17], InstructSpeech [271], Step-Audio [265]

Fig. 4. A taxonomy of controllable TTS from the perspective of control strategies.

the generated speech. Spark-TTS [267] offers both coarse-grained control (e.g., gender, speaking style) and fine-grained adjustments (e.g., precise pitch values, speaking rate). Users can modify specially designed tokens through prompts to achieve customized speech synthesis with optional reference speech as input.

Other methods enable control by adjusting continuous input signals. DiffStyleTTS [221] models prosody hierarchically, incorporating coarse-grained implicit style conditions (extracted via GST [19]) and fine-grained explicit features. Users can adjust guiding scale factors to control pitch, energy, duration, and prosodic styles. DrawSpeech [222] provides an intuitive way to manipulate pitch and energy. Users can sketch rough prosody contours (e.g., rising or falling patterns), which the system refines into detailed contours through a sketch-to-contour predictor. A diffusion model then generates expressive speech that aligns with the user's sketched prosody, offering precise control over vocal emphasis and intonation.

In addition, some approaches regulate speech attributes by modifying latent features rather than direct input signals. Cauliflow [186] controls speech rate and pausing by conditioning its flow-based duration model on user-defined parameters (r_s for speed, r_p for pause frequency) that adjust deviations from dataset averages during synthesis. DurIAN-E [195] allows users to manipulate expressive styles. It employs variance predictors for prosodic attributes and incorporates Style-Adaptive Instance Normalization (SAIN) layers, which dynamically adjust mel-spectrogram statistics based on predefined style embeddings. DiTTo-TTS [205], a diffusion transformer-based TTS model, removes dependencies on domain-specific features such as phonemes and durations while maintaining high performance. It controls speech rate by modifying the latent length predicted by a length predictor.

These methods show great potential in controlling speech attributes by adjusting input signals or latent variables. However, these methods are limited in expressive diversity, as they can only model a small set of pre-defined attributes.

Reference Speech Prompt. This paradigm aims to customize a new speaker's voice with just a few seconds of voice prompt. The architecture can be abstracted into two main components: a speaker encoder that processes the reference speech and outputs a speaker embedding, and a conditional TTS decoder that takes both text and speaker embedding as input to generate speech that matches the style of the reference prompt. MetaStyleSpeech [184] and StyleTTS [88] use adaptive normalization as a style conditioning method, enabling robust zero-shot performance. GenerSpeech [90] introduces a multilevel style adapter to improve zero-shot style transfer for out-of-domain custom voices. SC VALL-E [242] facilitates control over synthesized speech's emotions, speaking styles, and various acoustic features by incorporating style tokens and scale factors. ArtSpeech [210] revisits the sound production system by integrating articulatory representations into the TTS framework, improving the physical interpretability of articulation movements.

To enhance the learning of contextual information and address the challenge of limited voice data from the target speaker, CCSP [211] proposes a contrastive context-speech pretraining (CCSP) framework that learns cross-modal representations, combining both contextual text and speech expressions. DEX-TTS [209] separates styles into time-invariant and time-variant components, enabling the extraction of diverse styles from expressive reference speech. StyleTTS-ZS [214] leverages distilled time-varying style diffusion to capture diverse speaker identities and prosodies.

Some works also decouple timbre and style information from the reference speech, allowing more flexible control over the speaking style [87], [106], [193]. MegaTTS 2 [193] introduces an acoustic autoencoder that separately encodes prosody and timbre into the latent space, enabling the transfer of various speaking styles to the desired timbre. ControlSpeech [106] uses bidirectional attention and mask-based parallel decoding to capture codec representations in a discrete decoupling codec space, allowing independent control of timbre, style, and content in a zero-shot manner.

Natural Language Descriptions. Recent studies explore controlling speech style using natural language descriptions that include attributes such as pitch, gender, and emotion, making the process more user-friendly and interpretable. In this paradigm, several speech datasets with natural language descriptions [101], [106], [243] and associated prompt generation pipelines [102], [243], [273] have been proposed. Detailed information about these datasets will be discussed in Section V. PromptTTS [101] uses manually annotated text prompts to describe five speech attributes, including gender, pitch, speaking speed, volume, and emotion. InstructTTS [105] introduces a three-stage training procedure to capture semantic information from natural language style prompts and adds further annotation to the NLSpeech dataset's speech styles. PromptStyle [190] constructs a shared space for stylistic and semantic representations through a two-stage training process. TextrolSpeech [243] proposes an efficient prompt programming methodology and a multi-stage discrete style token-guided control framework, demonstrating strong in-context capabilities. NansyTTS [215] combines a TTS trained on the target language with a description control model trained on another language, which shares the same timbre and style representations to enable cross-lingual controllability.

Considering that not all details about voice variability can be described in the text prompt, PromptTTS++ [104] and PromptSpeaker [272] try to construct text prompts with more details. PromptTTS 2 [102] designs a variation network to capture voice variability not conveyed by text prompts. ControlSpeech [106] proposes the Style Mixture Semantic Density (SMSD) module, incorporating a noise perturbation mechanism to tackle the many-to-many problem in style control and enhance style diversity.

Other works also focus on improving controllability in additional aspects, such as the surrounding environment. Audiobox [201] introduces both description-based and example-based prompting, integrating speech and sound generation paradigms to independently control transcript, vocal, and other audio styles during speech generation. VoiceLDM [194] and AST-LDM [203] extend AudioLDM [224] to incorporate environmental context in TTS by adding a content prompt as a conditional input. Building on VoiceLDM, MS^{2}KU-VTTS [217] further expands the dimensions of environmental perception, enhancing the generation of immersive speech.

Instruction-Guided Control. The description-based TTS methods discussed above require splitting inputs into content and description prompts, which limits fine-grained control over speech and does not align with other AIGC models. VoxInstruct [103] proposes a new paradigm that extends traditional text-to-speech tasks into a general human instruction-to-speech task. Here, human instructions are freely written in natural language, encompassing both the spoken content and descriptive information about the speech. To enable automatic extraction of the synthesized speech content from raw text instructions, VoxInstruct uses speech semantic tokens as an intermediate representation, bridging the gap in current research by allowing the simultaneous use of both text description prompts and speech prompts for speech generation. CosyVoice [17] introduces supervised semantic tokens derived from ASR models via vector quantization, enabling precise text-speech alignment. It employs an LLM for token generation and flow matching for synthesis. Controllability is enhanced through instruction fine-tuning, allowing adjustments in speaker identity, emotion, speaking rate, pitch, and paralinguistic elements (e.g., laughter, breath sounds) via textual instruction.

Some speech LLMs can also synthesize speech content, but offer less controllability. These speech LLMs are agents that take user instructions and call expert TTS models to synthesize speech content. AudioGPT [274] is a multimodal LLM designed to process and generate audio data. It extends the capabilities of general-purpose LLMs by incorporating audio/speech understanding, synthesis (FastSpeech2 [76]), and style conversion [90] modules. SpeechGPT [34] integrates speech and text via discrete representations. It employs a three-stage training strategy for cross-modal alignment. For speech synthesis control, it uses HiFi-GAN with speaker embeddings and chain-of-modality instructions, enabling instruction-guided speech generation, e.g., one can input "Please read the sentence: Today is a beautiful day." FunAudioLLM [275] is an LLM enhancing voice interaction between humans and LLMs via SenseVoice [275] (multilingual speech recognition) and CosyVoice [17] (controllable speech synthesis). StepAudio [265] introduces a unified 130B speech-text model integrating understanding/generation, with a generative data engine enabling affordable voice cloning and a controllable TTS system (Step-Audio-TTS-3B [265]). Its speech synthesis achieves dynamic control through instruction-driven adjustments of dialects, emotions (anger/joy/sadness), singing styles, RAP vocals, and speaking rates.

V. Datasets and Evaluation Metrics

A. Datasets

Fully controllable text-to-speech (TTS) systems require large-scale datasets that exhibit extensive diversity and include fine-grained annotations. Such datasets provide essential information that enables TTS models to generate highly expressive speech with precise control over various speech attributes.

In this subsection, we classify speech datasets into three categories based on the type of annotations they provide: (1) tag-based datasets, which include structured attribute labels such as age and gender; (2) description-based datasets, which contain detailed textual descriptions of speech characteristics; and (3) dialogue datasets, which capture natural conversational speech. Below, we provide an overview of these datasets and discuss their importance in advancing TTS research.

Tag-based Datasets. Tag-based datasets consist of speech recordings annotated with predefined attribute tags that characterize various aspects of the speech signal [277]–[280], [282], [284]–[288], [290]. These attributes include pitch, energy, speed, age, gender, emotion, emphasis, accent, and topic, among others. Such datasets serve as a critical resource for developing TTS models with enhanced expressiveness and fine-grained control over speech generation. By incorporating attribute labels, models can be trained to adjust specific characteristics dynamically, enabling more personalized and context-aware speech synthesis.

Description-based Datasets. Description-based datasets go beyond structured labels by pairing speech samples with detailed textual descriptions that characterize various speech attributes, such as intonation, prosody, speaking style, and emotional nuances [101], [106], [243], [273], [291]. Unlike tag-based datasets, which provide predefined categorical labels, description-based datasets enable models to interpret nuanced, free-form textual prompts and generate speech that aligns with complex, context-dependent specifications. This

TABLE VI
A SUMMARY OF PUBLICLY AVAILABLE SPEECH DATASETS FOR CONTROLLABLE TTS.

Dataset	Hours (at least)	#Speakers (at least)	Labels											Lang	Release Time
			Pit.	Ene.	Spe.	Age	Gen.	Emo.	Emp.	Acc.	Top.	Des.	Dia.		
IEMOCAP [277]	12	10	✓	✓	✓		✓	✓						en	2008
RECOLA [278]	3.8	46						✓						fr	2013
RAVDESS [279]	/	24				✓		✓						en	2018
CMU-MOSEI [280]	65	1,000						✓						en	2018
Taskmaster-1 [281]	/	/											✓	en	2019
AISHELL-3 [282]	85	218				✓	✓			✓				zh	2020
Common Voice [283]	2,500	50,000				✓	✓			✓				multi	2020
ESD [284]	29	10						✓						en,zh	2021
GigaSpeech [285]	10,000	/									✓			en	2021
WenetSpeech [286]	10,000	/									✓			zh	2021
PromptSpeech [101]	/	/	✓	✓	✓		✓					✓		en	2022
MagicData-RAMC [287]	180	663									✓		✓	zh	2022
DailyTalk [288]	20	2						✓			✓		✓	en	2023
TextrolSpeech [243]	330	1,324	✓	✓	✓		✓	✓				✓		en	2023
CLESC [289]	<1	/	✓	✓	✓			✓						en	2024
VccmDataset [106]	330	1,324	✓	✓	✓		✓	✓				✓		en	2024
MSceneSpeech [290]	13	13									✓			zh	2024
Parler-TTS [273]	50,000	/	✓		✓		✓	✓		✓		✓		en	2024
SpeechCraft [291]	2,391	3,200	✓	✓	✓	✓	✓	✓	✓		✓	✓		en,zh	2024

Abbreviations: Pit(ch), Ene(rgy)=volume, Spe(ed), Gen(der), Emo(tion), Emp(hasis), Acc(ent), Top(ic), Des(cription), Dia(logue).

category of datasets plays a crucial role in training TTS systems that can respond to natural language descriptions, making them highly suitable for applications requiring fine-grained expressiveness, such as storytelling, audiobooks, and personalized voice assistants.

Dialogue Datasets. Dialogue datasets [281], [287], [288] capture multi-turn conversational speech between two or more speakers, focusing on the natural flow of human interaction, including turn-taking, contextual dependencies, and prosodic variations. These datasets are particularly valuable for training conversational TTS models that can generate dynamic and contextually appropriate speech for chatbots and interactive applications. By incorporating dialogue-specific characteristics such as speaker intent, pauses, and conversational nuances, models trained on these datasets can produce more natural and engaging interactions for real-world scenarios.

By leveraging these different categories of speech datasets, researchers can develop more advanced and flexible TTS models capable of generating speech that is not only intelligible but also expressive, context-sensitive, and highly controllable. We summarize publicly available datasets in Table VI.

B. Evaluation

The performance of controllable TTS often requires objective and subjective evaluation. We introduce common evaluation metrics in this subsection.

Objective Evaluation Metrics. Objective metrics offer automated and reproducible evaluations. Mel Cepstral Distortion (MCD) [292] measures the spectral distance between synthesized and reference speech, reflecting how closely the generated audio matches the target in terms of acoustic features. A lower MCD value indicates a higher similarity between synthesized and reference speech, meaning better speech synthesis quality. Typically, an MCD value below 4 suggests good quality, while values above 6 may indicate significant distortion. The MCD is computed as follows:

TABLE VII
COMMON OBJECTIVE AND SUBJECTIVE EVALUATION METRICS.

Metric	Type	Eval Target	GT Required
MCD [292]$\downarrow$	Objective	Acoustic similarity	✓
FDSD [293]$\downarrow$	Objective	Acoustic similarity	✓
WER [294]$\downarrow$	Objective	Intelligibility	✓
Cosine [295], [296]$\downarrow$	Objective	Speaker similarity	✓
PESQ [297]$\uparrow$	Objective	Perceptual quality	✓
SNR [298]$\uparrow$	Objective	Perceptual quality	✓
MOS [299]$\uparrow$	Subjective	Preference	
CMOS [300]$\uparrow$	Subjective	Preference	
AB Test	Subjective	Preference	
ABX Test	Subjective	Perceptual similarity	✓

GT: Ground truth, $\downarrow$: Lower is better, $\uparrow$: Higher is better.

$$MCD = \frac{10}{\ln 10} \cdot \sqrt{2\sum_{d=1}^{D}(c_d^{(syn)} - c_d^{(ref)})^2}, \qquad (3)$$

where $c_d^{(syn)}$ represents the d-th Mel Cepstral Coefficient (MCC) of the synthesized speech, $c_d^{(ref)}$ represents the d-th MCC of the reference speech, D is the number of MCC, and $\frac{10}{\ln 10} \approx 4.342$ is a constant factor that converts the logarithm to a decibel scale.

Fréchet DeepSpeech Distance (FDSD) [293] is another metric designed to evaluate the quality and naturalness of synthesized speech. It is inspired by the Fréchet Inception Distance (FID) [301] used in image generation but adapted to speech by leveraging a deep speech recognition model. FDSD measures the statistical distance between the distributions of real (reference) and synthesized speech in the feature space of a pretrained speech recognition model, such as Deep Speech [302]. By comparing the mean and covariance of the extracted feature representations, FDSD provides a perceptually relevant assessment of speech synthesis quality. A lower FDSD means the synthesized speech is more similar to real speech. FDSD can be computed as:

$$FDSD = ||\mu_s - \mu_r||^2 + \text{Tr}(\Sigma_s + \Sigma_r - 2(\Sigma_s \Sigma_r)^{1/2}), \quad (4)$$

where μ_s and Σ_s are the mean and covariance of the embeddings from the synthesized speech, μ_r and Σ_r are the mean and covariance of the embeddings from the real (reference) speech, $||\mu_s - \mu_r||^2$ represents the squared Euclidean distance between the means, $\text{Tr}(\cdot)$ denotes the trace of a matrix, and $(\Sigma_s \Sigma_r)^{1/2}$ is the geometric mean of the covariance matrices.

For intelligibility, the Word Error Rate (WER) [294] is used. It measures the difference between the recognized transcript and the reference transcript by computing the number of errors made in the transcription process. WER is computed as:

$$WER = \frac{S + D + I}{N}, \quad (5)$$

where S is the number of substitutions (wrong word in place of the correct word), D is the number of deletions (missed words), I is the number of insertions (extra words added), and N is the total number of words in the reference transcript.

Cosine similarity (on speaker embeddings) measures similarity between speaker embeddings of synthesized and reference speech. It can be used to evaluate zero-shot TTS (voice cloning) methods, where higher values indicate better speaker similarity. Given two speaker embeddings, $\mathbf{e_1}$ and $\mathbf{e_2}$, their cosine similarity is defined as:

$$CosSim(\mathbf{e_1}, \mathbf{e_2}) = \frac{\mathbf{e_1} \cdot \mathbf{e_2}}{\|\mathbf{e_1}\| \|\mathbf{e_2}\|}, \quad (6)$$

where speaker embeddings can be extracted from a pre-trained speaker embedding model (e.g., ECAPA-TDNN [295] and x-vectors [296]).

Perceptual Evaluation of Speech Quality (PESQ) [297] is another objective metric designed to evaluate speech quality by comparing degraded audio with a clean reference. It is widely used in telecommunications and speech synthesis. PESQ models human auditory perception, producing a score in the range $[-0.5, -4.5]$ that reflects intelligibility and distortion under various conditions, including noise or compression. PESQ involves complex perceptual modeling, its core components can be summarized as:

$$PESQ = a_0 + a_1 \cdot D_{frame} + a_2 \cdot D_{time}, \quad (7)$$

where D_{frame} is the frame-by-frame perceptual distortion, D_{time} is the time-domain distortion, and a_0, a_1, a_2 are regression coefficients. One can refer to [297] for details.

Signal-to-Noise Ratio (SNR) measures the ratio of signal power to noise power. A higher SNR indicates a cleaner signal with less noise, while a lower SNR suggests that noise is dominating the signal. However, in TTS, noise can come from different sources, such as artifacts from vocoders, neural network distortions, or background noise in dataset recordings. A direct computation of SNR in TTS requires a reference clean speech signal ($x[n]$), a synthesized (or noisy) speech signal ($y[n]$), and extracting the noise component ($e[n] = y[n] - x[n]$) from the synthesized signal. The SNR for TTS systems can be computed as:

$$SNR = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right), \quad (8)$$

where $P_{\text{signal}} = \frac{1}{N} \sum_{n=1}^{N} x[n]^2$ and $P_{\text{noise}} = \frac{1}{N} \sum_{n=1}^{N} e[n]^2$.

Subjective Evaluation Metrics. The Mean Opinion Score (MOS) [299] is the most commonly used subjective metric. In MOS evaluations, listeners rate various aspects, such as naturalness, expressiveness, quality, intelligibility, et al., of synthesized speech on a scale from 1 to 5, where higher scores indicate better quality. MOS captures human perception effectively but is expensive for large-scale evaluations.

Comparison Mean Opinion Score (CMOS) [300] further evaluates relative quality differences between two TTS audio samples. Participants listen to paired samples and rate their preference on a scale (e.g., -3 to +3, where negative values favor the first sample). CMOS is used to measure subtle improvements in TTS systems, complementing absolute MOS ratings. MOS and CMOS scores are computed as the average scores across all listeners:

$$MOS/CMOS = \frac{1}{N} \sum_{i=1}^{N} s_i, \quad (9)$$

where s_i is the score given by the i-th listener, and N is the number of listeners.

AB and ABX tests are also popular in evaluating TTS methods. An AB test involves presenting two versions of a synthesized speech (from different TTS models) to human listeners and asking them to choose which they prefer. The goal is to assess which model produces better-sounding speech based on certain criteria, such as naturalness, intelligibility, or clarity. In an ABX test, listeners compare two synthesized speech samples to a reference speech sample and determine which one is closer in terms of timbre, prosody, emotion, and other relevant features. ABX tests are widely used in evaluating zero-shot TTS methods. The AB/ABX test score for a model m is:

$$Score_{AB}/Score_{ABX} = \frac{N_m}{N}, \quad (10)$$

where N_m represents the number of listeners who prefer the speech synthesized by model m, and N denotes the total number of listeners.

Table VII summarizes widely used metrics for TTS.

VI. Challenges and Future Directions

In this section, we elaborate on current challenges for fully controllable TTS and discuss promising future directions.

A. Challenges

Controllable TTS aims to synthesize speech while allowing precise control over speech characteristics such as pitch, duration, energy, prosody, speaking style, and emotion. While significant progress has been made, achieving truly controllable TTS remains a complex task due to the multifaceted nature of human speech and the technical challenges in modeling and synthesizing it. In this section, we delve into the primary challenges and analyze their underlying reasons.

Controllability. A critical challenge in controllable TTS is determining what aspects of speech should be controlled and how to control speech characteristics at a specific granularity. Different applications require varying levels of control granularity. For instance, audiobook narration may need sentence-level control of emotion, while conversational AI like ChatGPT may require word or phoneme-level control over prosody. Moreover, the emotion, prosody, and other characteristics of human speech are often intricately intertwined and can manifest across varying levels of granularity. Additionally, achieving fine-grained control requires high-resolution annotations and sophisticated models capable of handling subtle variations without compromising synthesis quality.

Although some LLM-based TTS methods such as VoxInstruct [103] can control various aspects of speech through attribute descriptions, determining the appropriate level of granularity for control and devising methods to achieve precise control at a *specific granularity* or to enable *multiscale and fine-grained control* remains a significant challenge.

Feature Disentanglement and Representation. Achieving fully controllable TTS needs good feature disentanglement. Accurately extracting meaningful and disentangled speech features like pitch contours, energy patterns, emotion variation, and prosodic elements from training data is difficult. The reason is that speech features are interdependent and context-sensitive, making it hard to isolate specific attributes for control. For example, altering pitch often affects prosody, emotion, and naturalness to some extent. To tackle this, several methods [303]–[305] utilize pre-trained models for different speech recognition tasks (e.g., pitch, energy, and duration prediction, gender classification, age estimation, and speaker verification) to supervise feature extraction. For example, NaturalSpeech3 [14] factorizes speech into separate feature subspaces to capture different speech attributes.

However, these methods are limited to coarse or high-level feature disentanglement, leaving a significant gap in *fully disentangled control*. On the other hand, selecting *suitable representations* (e.g., continuous variables like mel-spectrograms or latent embeddings like tokens) for controllable attributes is non-trivial because representations must be both interpretable for humans and expressive enough for TTS models. For example, transformer-based models are good at processing discrete tokens, while GAN and diffusion-based models excel in modeling continuous representations.

Scarcity of Datasets. High-quality, diverse, and appropriately annotated datasets are essential for training controllable TTS systems. However, such datasets are scarce and costly to construct. In addition, training data must encompass a wide range of styles, emotions, accents, and prosodic variations to enable versatile control because limited diversity in datasets can restrict the model's ability to generalize across unseen styles or emotions. Although there are some large-scale datasets, such as LibriTTS [306], Gigaspeech [285], and TextrolSpeech [243], their diversity is still not enough for fully controllable TTS due to the lack of corpora for *diverse content and scenarios* such as comedies, thrillers, and cartoons. Constructing these large-scale datasets with rich diversity is expensive and time-consuming.

Another obstacle is that creating datasets with fine-grained, attribute-specific annotations is labor-intensive and costly. In addition, manual annotation of speech attributes requires expert knowledge and is prone to inconsistencies and errors, particularly for subjective attributes like emotion. Currently, most datasets provide only coarse labels, such as gender, age, or a limited range of emotions. While some datasets, such as SpeechCraft [291] and Parler-TTS [273], include natural language descriptions of speech attributes, no existing dataset offers *fine-grained variations and annotations* within the speech of the same speakers. Publicly available datasets for controllable TTS are summarized in Table VI.

Generalization Ability. The ability of a TTS system to generalize effectively is crucial for producing natural, high-quality speech across a wide range of conditions, such as unseen speakers, languages, and topics. However, achieving robust generalization remains a significant challenge for modern TTS methods due to various factors. *Zero-shot controllable* TTS [78], [92] aims to synthesize speech for unseen speakers with various speech customizations, such as emotion, using minimal reference audio, which can offer flexibility for personalized voice generation. However, it faces significant challenges, including capturing unique speaker characteristics from limited data, accurately reproducing prosody and style, and disentangling speaker identity from other attributes such as emotion and background noise.

Multilingual & low-resource generalization [253], [307] in TTS refers to the ability to synthesize natural and intelligible speech across multiple languages, including those not seen during training. This capability is essential for applications like cross-lingual communication, multilingual virtual assistants, and speech synthesis for low-resource languages [308]. Multilingual generalization still faces many challenges, such as linguistic diversity, mismatch, and the scarcity of data. Cross-lingual speaker generalization is another hurdle, as preserving speaker identity across languages can lead to artifacts.

Domain adaptation [309] in TTS refers to tailoring a pre-trained TTS model to generate speech for a specific domain or context, such as medical terminology and debate. One challenge is that many specialized domains lack sufficient high-quality annotated data for fine-tuning. In addition, adapting prosody, intonation, and speaking style to match domain-specific requirements such as comic dialogue is complex. Failing to capture domain-specific nuances can make speech sound unnatural or inconsistent with the target context.

Efficiency. Efficiency in controllable TTS systems is a critical requirement for practical applications, as these models aim to offer fine-grained control over various speech attributes such as prosody, emotion, style, and speaker identity. However, achieving such control often comes at the cost of increased computational complexity, larger model sizes, and longer inference times, creating significant challenges.

High latency is a major issue, as existing controllable TTS models [78], [101]–[103] often necessitate autoregressive processes to synthesize speech. The inference time for these models can range from several seconds to tens of seconds for a

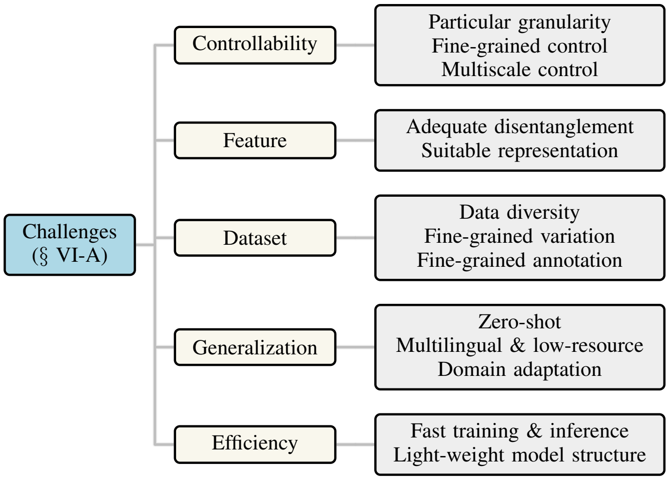


Fig. 5. A summary of current challenges for fully controllable TTS.

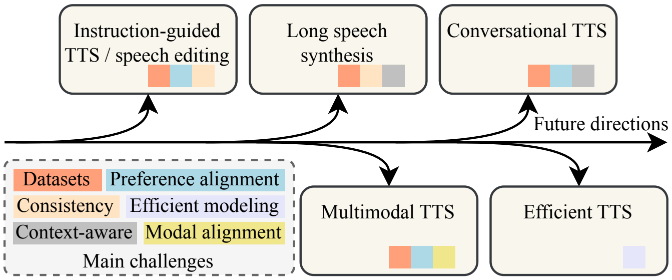


Fig. 6. Future directions and their top-3 challenges.

short audio clip (e.g., 5 to 10 seconds), and they do not support streaming synthesis. This can be particularly problematic for real-time applications like live broadcasting or interactive systems. Additionally, balancing granularity and efficiency presents a challenge. Finer controls require higher-resolution data and more precise modeling, which increases resource demands and results in *inefficient training and inference*.

Another major obstacle lies in the trade-off between model complexity and performance. SOTA controllable TTS systems often rely on LLMs with billions of parameters, which provide superior naturalness and expressiveness but demand significant computational resources. Simplifying these architectures can lead to quality degradation, including artifacts, unnatural prosody, or limited expressiveness. Therefore, designing *lightweight* controllable TTS models is significantly tricky.

B. Future Directions

In this survey, we conduct a comprehensive investigation and analysis of existing TTS methods, particularly on controllable TTS technologies. While these methods show great potential in applications, numerous limitations and challenges still need to be addressed. Hence, we are still in the early stage of controllable speech synthesis. Based on our observations, we outline several promising future directions as follows:

Instruction-Guided Fine-grained Speech Synthesis. Using natural language description to synthesize human speech with fine-grained control over various speech attributes is currently underexplored. Most of the existing works can only control a fixed number of attributes of the synthesized speech. Although a few works show great control of emotion, timbres, pitch, gender, and styles, e.g., VoxInstruct [103] and CosyVoice [17], they can frequently synthesize unwanted speech clips inconsistent with user instructions. Users often need to synthesize multiple times to get satisfactory speech.

Instruction-Guided Fine-grained Speech Editing. Speech or audio editing has been studied for a long time. However, existing methods usually train conditional models and adjust a fixed number of conditional inputs to modify the attributes of synthesized speech, thus lacking fine-grained manipulations [94], [95]. Therefore, how to learn disentangled speech representations for speech attributes while supporting instruction-guided editing is worthy of investigation.

Expressive Multi-modal Speech Synthesis. Synthesizing speech from multi-modal data such as texts, images, and videos is an appealing research topic due to its various applications in the industry, such as storytelling, filming, and gaming. Although there are several related works on this task [6], [24], [310], [311], they are still limited in extracting the required information from multimodal data. Particularly, synthesizing engaging speech and expressive voiceover for complex visual content sees great opportunities in the future.

Natural and Emotional Conversational TTS. Conversational TTS has been studied for several decades, but remained as cascaded systems for a long time, limiting its ability to generate natural and expressive speech. These systems are not context-aware, making the synthesized speech sound robotic. With the advent of LLMs, existing TTS technologies are directly introduced to synthesize speech using discrete speech tokens [33], [34]. However, context-aware conversational TTS with rich emotion and naturalness has not been well studied.

Zero-shot Long Speech Synthesis with Emotion Consistency. Zero-shot TTS is capable of voice cloning and speech style imitation without fine-tuning, making it practical in real scenarios [17], [78], [229]. However, synthesizing long speech with rich emotion and style variation in a zero-shot setting remains challenging due to the limited information in short reference audio clips. Addressing this issue will make a big step towards personalized zero-shot TTS.

Instruction-Guided Efficient TTS. Synthesizing speech with user instructions usually involves training language model-based codecs and bridge nets between different modalities, leading to much more computation overhead than previous TTS methods. The inference time is also relatively long, e.g., existing instruction-guided methods usually take tens of seconds to synthesize a short speech of less than 10 seconds [17], [104]. Therefore, efficient text and speech modeling is critical for instruction-guided TTS systems.

VII. Impacts of Controllable TTS

A. Applications

Controllable TTS enables fine-grained manipulation of speech attributes like pitch, emotion, and style, making it valuable across industries. For example, in virtual assistants and customer support, controllable TTS ensures context-aware responses, such as a calm tone for technical help or an enthusiastic pitch for promotions. In entertainment, it enhances voiceovers, audiobooks, and gaming characters by adjusting tone and delivery to match emotions and personalities. Education benefits from adaptive TTS, like slow, clear articulation for language learning or engaging narration for children.

For assistive technologies, controllable TTS empowers individuals with speech impairments to express emotions naturally. In content localization, it adapts speech to cultural preferences, ensuring a seamless experience for global audiences. Additionally, in human-computer interaction, it enables adaptive dialogue systems that adjust speech based on user mood or environment. By offering flexibility and expressiveness, controllable TTS enhances accessibility, personalization, and engagement across diverse applications.

B. Security Issues

One major security issue brought by controllable TTS is deepfakes. A deepfake is a type of synthetic media in which a person in an existing image, video, or audio recording is replaced with someone else's likeness or voice. This technology uses deep learning, particularly GANs [312], to create highly realistic but fabricated content. While deepfakes are most commonly associated with video manipulation, such as face swapping [313], they can also be applied to audio, enabling the creation of synthetic speech that mimics a specific person's voice, which is well known as voice cloning. Voice cloning, especially few-shot [314] and zero-shot TTS [78], [85], poses a significant threat to systems that rely on voice authentication, such as banking, customer service, and other identity verification processes, allowing attackers to impersonate individuals to gain unauthorized access to sensitive information or accounts.

Another issue is adversarial attacks [315]. Attackers can manipulate pitch, prosody, and phoneme duration to fool automatic speech recognition and speaker verification. It also poses risks in data poisoning [316], where adversarially modified samples degrade model performance. On the other hand, controllable TTS can strengthen adversarial defenses by generating diverse speech variations for adversarial training and counter-adversarial augmentation [317].

To address these concerns, it's essential to establish robust security protocols, consent-based regulations, and public awareness around voice cloning. Furthermore, advancements in detecting voice clones are equally important to help distinguish genuine voices from synthesized ones, protecting both individuals and organizations from potential misuse.

VIII. Conclusion

In this survey, we have first elaborated on the general pipeline for controllable TTS, followed by a glimpse of uncontrollable TTS methods from the perspective of local and global speech modeling. Then, we have comprehensively reviewed existing controllable methods from the perspectives of model architectures and control strategies. Popular datasets and commonly used evaluation metrics for controllable TTS were also summarized in this paper. Besides, the current challenges were deeply analyzed, and the promising future directions were also pointed out. To the best of our knowledge, this is the first comprehensive survey for controllable TTS.

References

[1] Wikipedia, "Speech synthesis." https://en.wikipedia.org/wiki/Speech_synthesis. Accessed: 2024-10-19.

[2] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, vol. 3. Springer Science & Business Media, 1997.

[3] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.

[4] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[5] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. Siri vs. Cortana vs. Google assistant: A comparison of speech-based natural user interfaces," in *Advances in Human Factors and Systems Interaction*, pp. 241–250, 2018.

[6] Y. Li, F. Yu, Y.-Q. Xu, E. Chang, and H.-Y. Shum, "Speech-driven cartoon animation with emotions," in *Proceedings of the Ninth ACM International Conference on Multimedia*, pp. 365–371, 2001.

[7] Y. Wang, W. Wang, W. Liang, and L.-F. Yu, "Comic-guided speech synthesis," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1–14, 2019.

[8] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey, *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[9] S. Roehling, B. MacDonald, and C. Watson, "Towards expressive speech synthesis in English on a robotic platform," in *Proceedings of the Australasian International Conference on Speech Science and Technology*, pp. 130–135, 2006.

[10] OpenAI, "Introducing ChatGPT." https://openai.com/index/chatgpt/. Accessed: 2024-10-22.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, "GPU computing," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 879–899, 2008.

[14] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, S. Zhao, T. Qin, F. Soong, and T.-Y. Liu, "NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, 2024.

[15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–10, 2019.

[16] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, *et al.*, "Deep Voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*, pp. 195–204, 2017.

[17] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, *et al.*, "CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[18] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6945–6949, 2019.

[19] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, pp. 5167–5176, 2018.

[20] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7254–7258, 2020.

[21] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *12th International Symposium on Chinese Spoken Language Processing*, pp. 1–5, 2021.

[22] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody

and timbre," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1693–1705, 2023.

[23] J. Wang, Z. Wang, X. Hu, X. Li, Q. Fang, and L. Liu, "Residual-guided personalized speech synthesis based on face image," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4743–4747, 2022.

[24] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image.," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1321–1325, 2020.

[25] J. Choi, J. Hong, and Y. M. Ro, "DiffV2S: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7812–7821, 2023.

[26] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[27] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[28] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, *et al.*, "DeepSeek LLM: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.

[29] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, *et al.*, "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[30] H. Hao, L. Zhou, S. Liu, J. Li, S. Hu, R. Wang, and F. Wei, "Boosting large language model for speech synthesis: An empirical study," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

[31] X. Gao, C. Zhang, Y. Chen, H. Zhang, and N. F. Chen, "Emo-DPO: Controllable emotional speech synthesis through direct preference optimization," *arXiv preprint arXiv:2409.10157*, 2024.

[32] P. Neekhara, S. Hussain, S. Ghosh, J. Li, R. Valle, R. Badlani, and B. Ginsburg, "Improving robustness of LLM-based speech synthesis by learning monotonic alignment," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 3425–3429, 2024.

[33] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "LLaMA-Omni: Seamless speech interaction with large language models," in *International Conference on Learning Representations*, pp. 1–18, 2025.

[34] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15757–15773, 2023.

[35] X. Zhang, X. Lyu, Z. Du, Q. Chen, D. Zhang, H. Hu, C. Tan, T. Zhao, Y. Wang, B. Zhang, *et al.*, "IntrinsicVoice: Empowering llms with intrinsic real-time voice interaction abilities," *arXiv preprint arXiv:2410.08035*, 2024.

[36] D. H. Klatt, "Review of text-to-speech conversion for english," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[37] T. Dutoit, "High-quality text-to-speech synthesis: An overview," *Journal of Electrical and Electronics Engineering Australia*, vol. 17, no. 1, pp. 25–36, 1997.

[38] A. Breen, "Speech synthesis models: a review," *Electronics & Communication Engineering Journal*, vol. 4, no. 1, pp. 19–31, 1992.

[39] J. P. Olive and M. Y. Liberman, "Text to speech—an overview," *The Journal of the Acoustical Society of America*, vol. 78, no. S1, pp. S6–S6, 1985.

[40] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[41] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[42] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[43] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.

[44] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: A review," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 5837–5880, 2023.

[45] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[46] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1355–1381, 2023.

[47] Z. Mu, X. Yang, and Y. Dong, "Review of end-to-end speech synthesis technology based on deep learning," *arXiv preprint arXiv:2104.09995*, 2021.

[48] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.

[49] Y. Tabet and M. Boughazi, "Speech synthesis techniques. a survey," in *International Workshop on Systems, Signal Processing and their Applications*, pp. 67–70, 2011.

[50] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI," *arXiv preprint arXiv:2303.13336*, 2023.

[51] L. R. Rabiner, "Digital-formant synthesizer for speech-synthesis studies," *The Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 822–828, 1968.

[52] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.

[53] D. W. Purcell and K. G. Munhall, "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 966–977, 2006.

[54] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, p. 971 – 995, 1980.

[55] J. Wouters and M. W. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 30–38, 2001.

[56] M. Bulut, S. S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer.," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1265–1268, 2002.

[57] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 781–784, 2001.

[58] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[59] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 373–376, 1996.

[60] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.

[61] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[62] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Transactions on Information and Systems*, vol. 92, no. 3, pp. 489–497, 2009.

[63] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis.," in *5th International Conference on Spoken Language Processing*, vol. 98, pp. 29–32, 1998.

[64] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *6th European Conference on Speech Communication and Technology*, pp. 2347–2350, 1999.

[65] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1315–1318, 2000.

[66] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[67] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[68] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1611–1614, 1997.

[69] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, 2006.

[70] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2461–2464, 2003.

[71] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 88, no. 3, pp. 502–509, 2005.

[72] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, "Emotion transplantation through adaptation in HMM-based speech synthesis," *Computer Speech & Language*, vol. 34, no. 1, pp. 292–307, 2015.

[73] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, *et al.*, "WaveNet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, vol. 12, p. 125, 2016.

[74] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Annual Conference of the International Speech Communication Association*, pp. 4006–4010, 2017.

[75] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.

[76] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, pp. 1–15, 2021.

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

[78] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "MaskGCT: Zero-shot text-to-speech with masked generative codec transformer," in *International Conference on Learning Representations*, pp. 1–24, 2025.

[79] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4475–4479, 2015.

[80] S.-F. Huang, C.-J. Lin, D.-R. Liu, Y.-C. Chen, and H.-y. Lee, "Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1558–1571, 2022.

[81] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "MultiSpeech: Multi-speaker text to speech with transformer," in *Annual Conference of the International Speech Communication Association*, pp. 4024–4028, 2020.

[82] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*, pp. 2709–2720, 2022.

[83] Y. Lei, S. Yang, X. Wang, and L. Xie, "MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[84] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5709–5713, 2021.

[85] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[86] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, sheng zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *International Conference on Learning Representations*, pp. 1–25, 2024.

[87] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, E. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and sheng zhao, "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *International Conference on Machine Learning*, pp. 1–19, 2024.

[88] Y. A. Li, C. Han, and N. Mesgarani, "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 1, pp. 283–296, 2025.

[89] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *37th Conference on Neural Information Processing Systems*, pp. 1–28, 2023.

[90] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "GenerSpeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Advances in Neural Information Processing Systems*, pp. 1–14, 2022.

[91] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, *et al.*, "Mega-TTS: Zero-shot text-to-speech at scale with intrinsic inductive bias," *arXiv preprint arXiv:2306.03509*, 2023.

[92] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6184–6188, 2020.

[93] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–12462, 2024.

[94] D. Tan, L. Deng, Y. T. Yeung, X. Jiang, X. Chen, and T. Lee, "EditSpeech: A text based speech editing system using partial inference and bidirectional fusion," in *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 626–633, 2021.

[95] J. Tae, H. Kim, and T. Kim, "EdiTTS: Score-based editing for controllable text-to-speech," in *Annual Conference of the International Speech Communication Association*, pp. 421–425, 2022.

[96] S. Seshadri, T. Raitio, D. Castellani, and J. Li, "Emphasis control for parallel neural TTS," in *Annual Conference of the International Speech Communication Association*, pp. 3378–3382, 2022.

[97] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[98] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[99] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[100] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "PaLM: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[101] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[102] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, *et al.*, "PromptTTS 2: Describing and generating voices with text prompt," in *The Twelfth International Conference on Learning Representations*, 2023.

[103] Y. Zhou, X. Qin, Z. Jin, S. Zhou, S. Lei, S. Zhou, Z. Wu, and J. Jia, "VoxInstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 554–563, 2024.

[104] R. Shimizu, R. Yamamoto, M. Kawamura, Y. Shirahata, H. Doi, T. Komatsu, and K. Tachibana, "PromptTTS++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12672–12676, 2024.

[105] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2913–2925, 2024.

[106] S. Ji, J. Zuo, W. Wang, M. Fang, S. Zheng, Q. Chen, Z. Jiang, H. Huang, Z. Wang, X. Cheng, *et al.*, "ControlSpeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec," *arXiv preprint arXiv:2406.01205*, 2024.

[107] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 137–140, 1992.

[108] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[109] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "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*, vol. 27, no. 3-4, pp. 187–207, 1999.

[110] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7962–7966, 2013.

[111] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks.," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1964–1968, 2014.

[112] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A denoising diffusion model for text-to-speech," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 3605–3609, 2021.

[113] X. Chen, L. Xu, Z. Liu, M. Sun, and H.-B. Luan, "Joint learning of character and word embeddings," in *Twenty-fourth International Joint Conference on Artificial Intelligence*, pp. 1236–1242, 2015.

[114] F. Almeida and G. Xexéo, "Word embeddings: A survey," *arXiv preprint arXiv:1901.09069*, 2019.

[115] 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. 17022–17033, 2020.

[116] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *6th ISCA Workshop on Speech Synthesis*, p. 294 – 299, 2007.

[117] T. Nose and T. Kobayashi, "An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model," *Speech Communication*, vol. 55, no. 2, pp. 347–357, 2013.

[118] Y. Nishigaki, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Prosody-controllable HMM-based speech synthesis using speech input," *Proceedings of International Workshop on Machine Learning in Spoken Language Processing*, 2015.

[119] S. Vasquez and M. Lewis, "MelNet: A generative model for audio in the frequency domain," *arXiv preprint arXiv:1906.01083*, 2019.

[120] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[121] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in *International Conference on Learning Representations*, 2018.

[122] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *International Conference on Machine Learning*, pp. 7586–7598, 2020.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–11, 2017.

[124] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6706–6713, 2019.

[125] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[126] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, "Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13198–13206, 2021.

[127] S. Ma, D. Mcduff, and Y. Song, "Neural TTS stylization with adversarial and collaborative games," in *International Conference on Learning Representations*, 2018.

[128] H. Guo, F. K. Soong, L. He, and L. Xie, "A new GAN-based end-to-end TTS training algorithm," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1288–1292, 2019.

[129] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2018.

[130] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, pp. 8599–8608, 2021.

[131] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7209–7213, 2020.

[132] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[133] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[134] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[135] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *International Conference on Learning Representations*, 2017.

[136] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*, pp. 2410–2419, 2018.

[137] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5891–5895, 2019.

[138] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, *et al.*, "DurIAN: Duration informed attention network for speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2027–2031, 2020.

[139] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*, pp. 3918–3926, 2018.

[140] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *IEEE international conference on acoustics, speech and signal processing*, pp. 2251–2255, 2018.

[141] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*, 2018.

[142] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2019.

[143] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6199–6203, 2020.

[144] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14920–14921, 2019.

[145] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[146] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

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

[148] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "FastDiff: A fast conditional diffusion model for high-quality speech synthesis," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 4157–4163, 2022.

[149] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations*, pp. 1–17, 2021.

[150] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, pp. 1–15, 2021.

[151] S.-g. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu, "PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *International Conference on Learning Representations*, 2022.

[152] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting oversmoothness in text to speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8197–8213, 2022.

[153] S. Kim, K. Shih, J. F. Santos, E. Bakhturina, M. Desta, R. Valle, S. Yoon, B. Catanzaro, *et al.*, "P-Flow: a fast and data-efficient zero-shot TTS through speech prompting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74213–74228, 2024.

[154] Y. Guo, C. Du, Z. Ma, X. Chen, and K. Yu, "VoiceFlow: Efficient text-to-speech with rectified flow matching," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11121–11125, 2024.

[155] S.-H. Lee, H.-Y. Choi, and S.-W. Lee, "PeriodWave: Multi-period flow matching for high-fidelity waveform generation," *arXiv preprint arXiv:2408.07547*, 2024.

[156] W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A compact flow-based model for raw audio," in *International Conference on Machine Learning*, pp. 7706–7716, 2020.

[157] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," in *International Conference on Machine Learning*, pp. 3370–3378, 2019.

[158] H. Guo, F. Xie, X. Wu, F. K. Soong, and H. Meng, "MSMC-TTS: Multi-stage multi-codebook VQ-VAE based neural TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1811–1824, 2023.

[159] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, pp. 5530–5540, 2021.

[160] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *5th International Conference on Learning Representations, Workshop Track Proceedings*, 2017.

[161] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2019.

[162] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations*, 2021.

[163] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International Conference on Machine Learning*, pp. 1530–1538, 2015.

[164] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020.

[165] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[166] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[167] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, pp. 28492–28518, 2023.

[168] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, pp. 1298–1312, 2022.

[169] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[170] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[171] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

[172] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "HiFi-Codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[173] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[174] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[175] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[176] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li, Z. Zhang, X. Yang, R. Huang, Y. Jiang, Q. Chen, S. Zheng, and Z. Zhao, "WavTokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling," in *The Thirteenth International Conference on Learning Representations*, 2025.

[177] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[178] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan, "WaveGrad 2: Iterative refinement for text-to-speech synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association*, 2021.

[179] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3617–3621, 2019.

[180] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[181] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5709–5713, 2021.

[182] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive text-to-speech using style tag," *Annual Conference of the International Speech Communication Association*, pp. 4663–4667, 2021.

[183] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. De Oliveira, A. C. Junior, A. d. S. Soares, S. M. Aluisio, and M. A. Ponti, "SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model," *arXiv preprint arXiv:2104.05557*, 2021.

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

[185] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "DelightfulTTS: The Microsoft speech synthesis system for blizzard challenge 2021," *arXiv preprint arXiv:2110.12612*, 2021.

[186] A. Abbas, T. Merritt, A. Moinet, S. Karlapati, E. Muszynska, S. Slangen, E. Gatti, and T. Drugman, "Expressive, variable, and controllable duration modelling in TTS," *arXiv preprint arXiv:2206.14165*, 2022.

[187] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel WaveNet," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6044–6048, 2021.

[188] Z. Liu, Q. Tian, C. Hu, X. Liu, M. Wu, Y. Wang, H. Zhao, and Y. Wang, "Controllable and lossless non-autoregressive end-to-end text-to-speech," *arXiv preprint arXiv:2207.06088*, 2022.

[189] M. Kang, D. Min, and S. J. Hwang, “Grad-StyleSpeech: Any-speaker adaptive text-to-speech synthesis with diffusion models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[190] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, and L. Xie, “PromptStyle: Controllable style transfer for text-to-speech with natural language descriptions,” *arXiv preprint arXiv:2305.19522*, 2023.

[191] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, “istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6207–6211, 2022.

[192] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, *et al.*, “Voicebox: Text-guided multilingual universal speech generation at scale,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[193] Z. Jiang, J. Liu, Y. Ren, J. He, Z. Ye, S. Ji, Q. Yang, C. Zhang, P. Wei, C. Wang, *et al.*, “Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis,” in *The Twelfth International Conference on Learning Representations*, 2024.

[194] Y. Lee, I. Yeon, J. Nam, and J. S. Chung, “VoiceLDM: Text-to-speech with environmental context,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12566–12571, 2024.

[195] Y. Gu, Y. Bian, G. Lei, C. Weng, and D. Su, “DurIAN-E: Duration informed attention network for expressive text-to-speech synthesis,” *arXiv preprint arXiv:2309.12792*, 2023.

[196] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *The Eleventh International Conference on Learning Representations*, 2023.

[197] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, “Generative pre-training for speech with flow matching,” in *The Twelfth International Conference on Learning Representations*, 2024.

[198] S. Kim, K. Shih, J. F. Santos, E. Bakhturina, M. Desta, R. Valle, S. Yoon, B. Catanzaro, *et al.*, “P-Flow: a fast and data-efficient zero-shot TTS through speech prompting,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[199] Y. Gao, N. Morioka, Y. Zhang, and N. Chen, “E3 TTS: Easy end-to-end diffusion-based text to speech,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8, 2023.

[200] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, “HierSpeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis,” *arXiv preprint arXiv:2311.12454*, 2023.

[201] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, *et al.*, “Audiobox: Unified audio generation with natural language prompts,” *arXiv preprint arXiv:2312.15821*, 2023.

[202] Z. Ye, Z. Ju, H. Liu, X. Tan, J. Chen, Y. Lu, P. Sun, J. Pan, W. Bian, S. He, *et al.*, “FlashSpeech: Efficient zero-shot speech synthesis,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6998–7007, 2024.

[203] M. Kim, S.-W. Chung, Y. Ji, H.-G. Kang, and M.-S. Choi, “Speak in the Scene: Diffusion-based acoustic scene transfer toward immersive speech generation,” in *Annual Conference of the International Speech Communication Association*, pp. 4883–4887, 2024.

[204] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, “SimpleSpeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 4398–4402, 2024.

[205] K. Lee, D. W. Kim, J. Kim, and J. Cho, “DiTTo-TTS: Efficient and scalable zero-shot text-to-speech with diffusion transformer,” *arXiv preprint arXiv:2406.11427*, 2024.

[206] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, *et al.*, “E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS,” in *IEEE Spoken Language Technology Workshop*, pp. 682–689, 2024.

[207] S. Ji, Z. Jiang, H. Wang, J. Zuo, and Z. Zhao, “MobileSpeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech,” in *The 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 13588–13600, 2024.

[208] H. Siuzdak, “Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis,” in *The Twelfth International Conference on Learning Representations*, 2024.

[209] H. J. Park, J. S. Kim, W. Shin, and S. W. Han, “DEX-TTS: Diffusion-based expressive text-to-speech with style modeling on time variability,” *arXiv preprint arXiv:2406.19135*, 2024.

[210] Z. Wang, Y. Wang, M. Li, and H. Huang, “ArtSpeech: Adaptive text-to-speech synthesis with articulatory representations,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 535–544, 2024.

[211] Y. Xiao, X. Wang, X. Tan, L. He, X. Zhu, S. Zhao, and T. Lee, “Contrastive context-speech pretraining for expressive text-to-speech synthesis,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2099–2107, 2024.

[212] D. Yang, R. Huang, Y. Wang, H. Guo, D. Chong, S. Liu, X. Wu, and H. Meng, “SimpleSpeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models,” *arXiv preprint arXiv:2408.13893*, 2024.

[213] Z. Liu, S. Wang, P. Zhu, M. Bi, and H. Li, “E1 TTS: Simple and fast non-autoregressive TTS,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

[214] Y. A. Li, X. Jiang, C. Han, and N. Mesgarani, “StyleTTS-ZS: Efficient high-quality zero-shot text-to-speech synthesis with distilled time-varying style diffusion,” *arXiv preprint arXiv:2409.10058*, 2024.

[215] R. Yamamoto, Y. Shirahata, M. Kawamura, and K. Tachibana, “Description-based controllable text-to-speech with cross-lingual voice control,” *arXiv preprint arXiv:2409.17452*, 2024.

[216] N. Park, H. Kim, C. H. Lee, J. Choi, J. Yeom, and S. Yoon, “NanoVoice: Efficient speaker-adaptive text-to-speech for multiple speakers,” *arXiv preprint arXiv:2409.15760*, 2024.

[217] S. He, R. Liu, and H. Li, “Multi-source spatial knowledge understanding for immersive visual text-to-speech,” *arXiv preprint arXiv:2410.14101*, 2024.

[218] D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee, “EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector,” *arXiv preprint arXiv:2411.02625*, 2024.

[219] G. Cong, J. Pan, L. Li, Y. Qi, Y. Peng, A. v. d. Hengel, J. Yang, and Q. Huang, “EmoDubber: Towards high quality and emotion controllable movie dubbing,” *arXiv preprint arXiv:2412.08988*, 2024.

[220] S. Inoue, K. Zhou, S. Wang, and H. Li, “Hierarchical control of emotion rendering in speech synthesis,” *arXiv preprint arXiv:2412.12498*, 2024.

[221] J. Liu, Z. Liu, Y. Hu, Y. Gao, S. Zhang, and Z. Ling, “DiffStyleTTS: Diffusion-based hierarchical prosody modeling for text-to-speech with diverse and controllable styles,” in *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5265–5272, 2025.

[222] W. Chen, S. Yang, G. Li, and X. Wu, “DrawSpeech: Expressive speech synthesis using prosodic sketches as control conditions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

[223] S. Zhang, A. Mehrish, Y. Li, and S. Poria, “PROEMO: Prompt-driven text-to-speech synthesis based on emotion and intensity control,” *arXiv preprint arXiv:2501.06276*, 2025.

[224] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, “AudioLDM: text-to-audio generation with latent diffusion models,” in *Proceedings of the 40th International Conference on Machine Learning*, pp. 21450–21474, 2023.

[225] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, “AudioLDM 2: Learning holistic audio generation with self-supervised pretraining,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[226] 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,” in *International Conference on Machine Learning*, pp. 13916–13932, 2023.

[227] X. Zhu, W. Tian, X. Wang, L. He, X. Wang, S. Zhao, and L. Xie, “CosyAudio: Improving audio generation with confidence scores and synthetic captions,” *arXiv preprint arXiv:2501.16761*, 2025.

[228] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *The Eleventh International Conference on Learning Representations*, 2023.

[229] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, “F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2024.

[230] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, “ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

[231] W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang, “Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[232] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and B. Freeman, "Improved distribution matching distillation for fast image synthesis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 47455–47487, 2024.

[233] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

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

[235] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *IEEE Spoken Language Technology Workshop*, pp. 595–602, IEEE, 2018.

[236] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv preprint arXiv:2005.05957*, 2020.

[237] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.

[238] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[239] R. Huang, C. Zhang, Y. Wang, D. Yang, L. Liu, Z. Ye, Z. Jiang, C. Weng, Z. Zhao, and D. Yu, "Make-A-Voice: Unified voice synthesis with discrete representation," *arXiv preprint arXiv:2305.19269*, 2023.

[240] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[241] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Annual Conference of the International Speech Communication Association*, pp. 2207–2211, 2021.

[242] D. Kim, S. Hong, and Y.-H. Choi, "SC VALL-E: Style-controllable zero-shot text to speech synthesizer," *arXiv preprint arXiv:2307.10550*, 2023.

[243] S. Ji, J. Zuo, M. Fang, Z. Jiang, F. Chen, X. Duan, B. Huai, and Z. Zhao, "TextrolSpeech: A text style control speech corpus with codec language text-to-speech models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10301–10305, 2024.

[244] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu, *et al.*, "UniAudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[245] Y. Song, Z. Chen, X. Wang, Z. Ma, and X. Chen, "ELLA-V: Stable neural codec language modeling with alignment-guided sequence reordering," *arXiv preprint arXiv:2401.07333*, 2024.

[246] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski, *et al.*, "BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[247] J. Kim, K. Lee, S. Chung, and J. Cho, "CLaM-TTS: Improving neural codec language model for zero-shot text-to-speech," *arXiv preprint arXiv:2404.02781*, 2024.

[248] D. Xin, X. Tan, K. Shen, Z. Ju, D. Yang, Y. Wang, S. Takamichi, H. Saruwatari, S. Liu, J. Li, *et al.*, "RALL-E: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis," *arXiv preprint arXiv:2404.03204*, 2024.

[249] Z. Liu, S. Wang, S. Inoue, Q. Bai, and H. Li, "Autoregressive diffusion transformer for text-to-speech synthesis," *arXiv preprint arXiv:2406.05551*, 2024.

[250] B. Han, L. Zhou, S. Liu, S. Chen, L. Meng, Y. Qian, Y. Liu, S. Zhao, J. Li, and F. Wei, "VALL-E R: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment," *arXiv preprint arXiv:2406.07855*, 2024.

[251] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[252] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, *et al.*, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[253] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Conference of the International Speech Communication Association*, pp. 4978–4982, 2024.

[254] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu, *et al.*, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[255] H.-H. Guo, K. Liu, F.-Y. Shen, Y.-C. Wu, F.-L. Xie, K. Xie, and K.-T. Xu, "FireRedTTS: A foundation text-to-speech framework for industry-level generative speech applications," *arXiv preprint arXiv:2409.03283*, 2024.

[256] H. Guo, F. Xie, D. Yang, X. Wu, and H. Meng, "Speaking from coarse to fine: Improving neural codec language model via multi-scale speech coding and generation," *arXiv preprint arXiv:2409.11630*, 2024.

[257] S. Chen, Y. Feng, L. He, T. He, W. He, Y. Hu, B. Lin, Y. Lin, Y. Pan, P. Tan, *et al.*, "Takin: A cohort of superior quality zero-shot speech generation models," *arXiv preprint arXiv:2409.12139*, 2024.

[258] Y. Nishimura, T. Hirose, M. Ohi, H. Nakayama, and N. Inoue, "HALL-E: Hierarchical neural codec language model for minute-long zero-shot text-to-speech synthesis," *arXiv preprint arXiv:2410.04380*, 2024.

[259] S. Liao, Y. Wang, T. Li, Y. Cheng, R. Zhang, R. Zhou, and Y. Xing, "Fish-Speech: Leveraging large language models for advanced multilingual text-to-speech synthesis," *arXiv preprint arXiv:2411.01156*, 2024.

[260] W. Chen, Z. Ma, R. Yan, Y. Liang, X. Li, R. Xu, Z. Niu, Y. Zhu, Y. Yang, Z. Liu, *et al.*, "SLAM-Omni: Timbre-controllable voice interaction system with single-stage training," *arXiv preprint arXiv:2412.15649*, 2024.

[261] Y. Yang, Z. Ma, S. Liu, J. Li, H. Wang, L. Meng, H. Sun, Y. Liang, R. Xu, Y. Hu, *et al.*, "Interleaved speech-text language models are simple streaming text to speech synthesizers," *arXiv preprint arXiv:2412.16102*, 2024.

[262] X. Zhu, W. Tian, and L. Xie, "Autoregressive speech synthesis with next-distribution prediction," *arXiv preprint arXiv:2412.16846*, 2024.

[263] Y.-X. Lu, H.-P. Du, Z.-Y. Sheng, Y. Ai, and Z.-H. Ling, "Incremental disentanglement for environment-aware zero-shot text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

[264] H. Li, Y. Li, X. Wang, J. Hu, Q. Xie, S. Yang, and L. Xie, "FleSpeech: Flexibly controllable speech generation with various prompts," *arXiv preprint arXiv:2501.04644*, 2025.

[265] A. Huang, B. Wu, B. Wang, C. Yan, C. Hu, C. Feng, F. Tian, F. Shen, J. Li, M. Chen, *et al.*, "Step-Audio: Unified understanding and generation in intelligent speech interaction," *arXiv preprint arXiv:2502.11946*, 2025.

[266] X. Zhang, X. Zhang, K. Peng, Z. Tang, V. Manohar, Y. Liu, J. Hwang, D. Li, Y. Wang, J. Chan, Y. Huang, Z. Wu, and M. Ma, "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement," in *The Thirteenth International Conference on Learning Representations*, 2025.

[267] X. Wang, M. Jiang, Z. Ma, Z. Zhang, S. Liu, L. Li, Z. Liang, Q. Zheng, R. Wang, X. Feng, *et al.*, "Spark-TTS: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens," *arXiv preprint arXiv:2503.01710*, 2025.

[268] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, "HMM-based emotional speech synthesis using average emotion model," in *Proceedings of 5th Chinese Spoken Language Processing*, pp. 233–240, Springer, 2006.

[269] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, pp. 1–19, 2023.

[270] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[271] R. Huang, R. Hu, Y. Wang, Z. Wang, X. Cheng, Z. Jiang, Z. Ye, D. Yang, L. Liu, P. Gao, and Z. Zhao, "InstructSpeech: Following speech editing instructions via large language models," in *Forty-first International Conference on Machine Learning*, 2024.

[272] Y. Zhang, G. Liu, Y. Lei, Y. Chen, H. Yin, L. Xie, and Z. Li, "Promptspeaker: Speaker generation based on text descriptions," in *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–7, IEEE, 2023.

[273] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint arXiv:2402.01912*, 2024.

[274] 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," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 23802–23804, 2024.

[275] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu, *et al.*, "FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs," *arXiv preprint arXiv:2407.04051*, 2024.

[276] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728–53741, 2023.

[277] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[278] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–8, 2013.

[279] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS One*, vol. 13, no. 5, p. e0196391, 2018.

[280] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

[281] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, A. Cedilnik, and K.-Y. Kim, "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.

[282] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus," in *Conference of the International Speech Communication Association*, pp. 2756–2760, 2021.

[283] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, 2020.

[284] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[285] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Annual Conference of the International Speech Communication Association*, pp. 3670–3674, 2021.

[286] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, *et al.*, "WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6182–6186, 2022.

[287] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, *et al.*, "Open Source MagicData-RAMC: A rich annotated mandarin conversational (RAMC) speech dataset," *arXiv preprint arXiv:2203.16844*, 2022.

[288] K. Lee, K. Park, and D. Kim, "DailyTalk: Spoken dialogue dataset for conversational text-to-speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[289] Toloka, "Crowd labeled emotions and speech characteristics." https://huggingface.co/datasets/toloka/CLESC, 2024. Accessed: 2024-03-23.

[290] Q. Yang, J. Zuo, Z. Su, Z. Jiang, M. Li, Z. Zhao, F. Chen, Z. Wang, and B. Huai, "MSceneSpeech: A multi-scene speech dataset for expressive speech synthesis," in *Annual Conference of the International Speech Communication Association 2024*, pp. 1845–1849, 2024.

[291] Z. Jin, J. Jia, Q. Wang, K. Li, S. Zhou, S. Zhou, X. Qin, and Z. Wu, "SpeechCraft: A fine-grained expressive speech dataset with natural language description," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1255–1264, 2024.

[292] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion.," in *Spoken Language Technologies for Under-Resourced Languages*, pp. 63–68, 2008.

[293] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2020.

[294] Wikipedia, "Word error rate." https://en.wikipedia.org/wiki/Word_error_rate, 2024. Accessed: 2024-12-07.

[295] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *21st Annual Conference of the International Speech Communication Association*, pp. 3830–3834, 2020.

[296] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333, 2018.

[297] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, pp. 749–752, 2001.

[298] Wikipedia, "Signal-to-noise ration." https://en.wikipedia.org/wiki/Signal-to-noise_ratio, 2025. Accessed: 2025-03-25.

[299] Wikipedia, "Mean opinion score." https://en.wikipedia.org/wiki/Mean_opinion_score. Accessed: 2024-12-07.

[300] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*, pp. 623–654, Springer, 2011.

[301] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[302] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep Speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[303] X. An, F. K. Soong, and L. Xie, "Disentangling style and speaker attributes for TTS style transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 646–658, 2022.

[304] W. Wang, Y. Song, and S. Jha, "Generalizable zero-shot speaker adaptive speech synthesis with disentangled representations," in *Annual Conference of the International Speech Communication Association 2023*, pp. 4454–4458, 2023.

[305] X. An, F. K. Soong, S. Yang, and L. Xie, "Effective and direct control of neural TTS prosody by removing interactions between different attributes," *Neural Networks*, vol. 143, pp. 250–260, 2021.

[306] 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," in *Annual Conference of the International Speech Communication Association*, pp. 1526–1530, 2019.

[307] H. Cho, W. Jung, J. Lee, and S. H. Woo, "SANE-TTS: Stable and natural end-to-end multilingual text-to-speech," in *Annual Conference of the International Speech Communication Association 2022*, pp. 1–5, 2022.

[308] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint arXiv:2006.07264*, 2020.

[309] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in Data Science and Information Engineering*, pp. 877–894, 2021.

[310] Y. Rong and L. Liu, "Seeing your speech style: A novel zero-shot identity-disentanglement face-based voice conversion," *arXiv preprint arXiv:2409.00700*, 2024.

[311] J. Lu, B. Sisman, R. Liu, M. Zhang, and H. Li, "VisualTTS: TTS with accurate lip-speech synchronization for automatic voice over," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8032–8036, 2022.

[312] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259–6276, 2022.

[313] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7184–7193, 2019.

[314] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[315] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[316] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[317] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.