A Survey of Text-to-Speech Systems: Architectures, Techniques, and Evaluation

Abstract

Text-to-Speech (TTS) systems have become integral to enhancing human-computer interaction, offering natural and intelligible speech synthesis. This survey explores the evolution of TTS technologies from parametric models to advanced neural architectures, highlighting the transformative impact of deep learning. Key advancements include neural vocoders, which improve speech quality, and innovative architectures like Tacotron and WaveNet, which enable more natural speech synthesis. The paper discusses the challenges of prosody modeling, multilingual capabilities, and real-time processing, emphasizing the importance of efficient TTS systems for diverse applications. Techniques for voice conversion and zero-shot TTS are examined, showcasing the ability to generate speech without prior exposure to a speaker's voice. Evaluation metrics, both objective and subjective, are reviewed to assess TTS effectiveness and naturalness. The integration of natural language processing techniques further enhances TTS systems, enabling more expressive and contextually appropriate speech synthesis. Future directions include optimizing data-efficient models, enhancing global applications, and exploring new methodologies for voice conversion and multilingual TTS. This comprehensive survey underscores the potential of TTS technologies to promote inclusivity and facilitate natural human-computer interactions across diverse linguistic and cultural contexts.

1 Introduction

1.1 Role of TTS in Human-Computer Interaction

Text-to-Speech (TTS) systems are crucial for facilitating human-computer interactions by generating natural and intelligible speech, significantly enhancing accessibility and user experience [1]. In low-resource environments, TTS can improve communication through scalable speech synthesis solutions [2]. Traditional TTS systems, often limited to controlled corpora, face challenges in synthesizing human-level speech from spontaneous data [3].

TTS systems also play a vital role in cross-speaker style transfer, addressing the challenge of disentangling speaker and style information in audio [4]. This capability is essential for developing expressive TTS systems that can replicate diverse speaking styles, thus enhancing user engagement. Evaluating and minimizing the discrepancies between real and synthetic speech distributions is critical for improving the perceived naturalness and user satisfaction of TTS outputs [5].

The integration of TTS into various applications, including virtual assistants and interactive storytelling, underscores its significance in modern technology. By employing advanced deep learning techniques, TTS systems generate synthetic voices that closely mimic human speech, improving human-computer interaction. Developers can select from various TTS technologies—such as concatenative, formant synthesis, and statistical parametric methods—based on factors like voice naturalness and system complexity. Innovations like neural and hybrid TTS continue to evolve, addressing

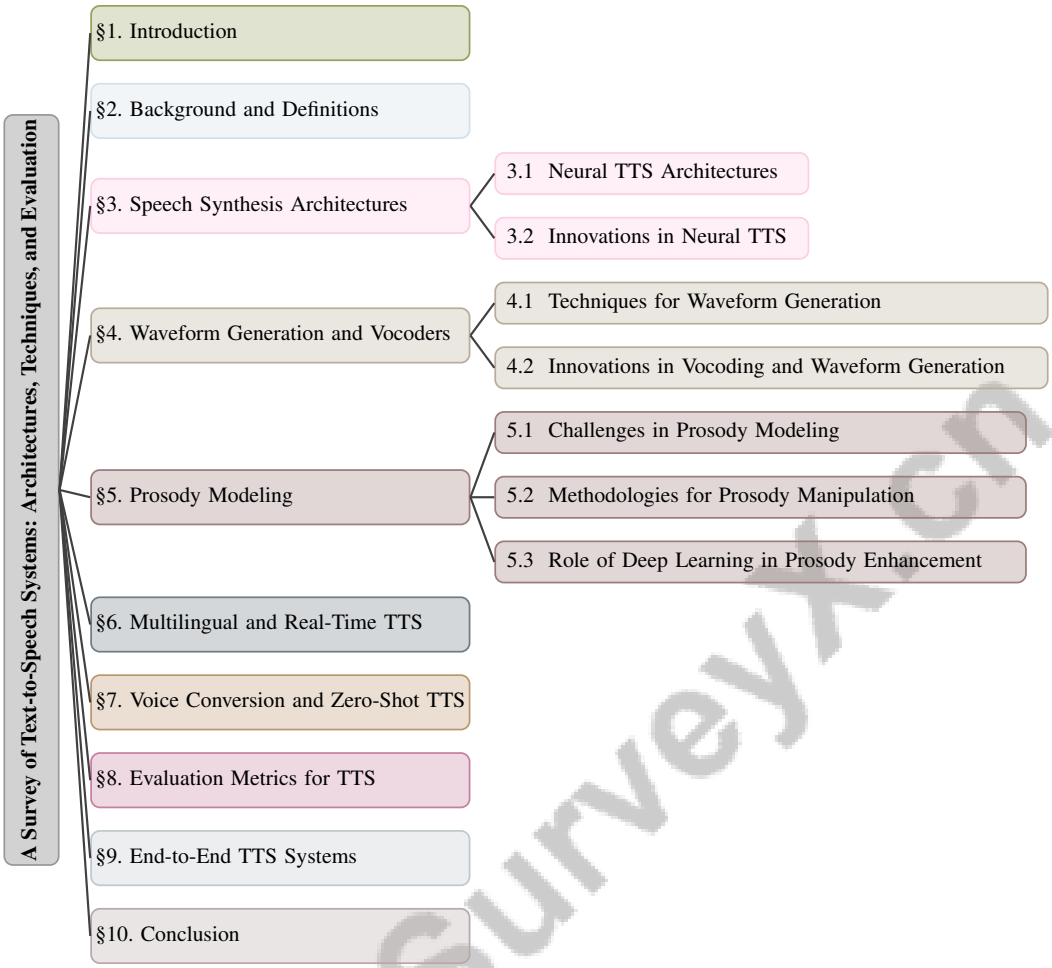


Figure 1: chapter structure

the limitations of current models, including issues of explainability and controllability over style modification, thereby promising more expressive and contextually relevant speech synthesis [6, 7].

1.2 Evolution of TTS Technologies

The evolution of TTS technologies has transitioned from early parametric models to sophisticated neural network-based systems, each marked by significant advancements in speech naturalness. Early TTS systems relied on rule-based algorithms and extensive linguistic knowledge, often producing outputs perceived as robotic and unnatural [6]. These systems were limited by their reliance on handcrafted features, restricting their adaptability and expressiveness.

The emergence of neural TTS systems has transformed the field. Deep learning techniques, particularly deep neural networks (DNNs), enable the modeling of speech in a more natural and flexible manner, allowing for the synthesis of human-like speech and overcoming the constraints of traditional parametric approaches. Neural TTS has shown particular promise in low-resource languages, demonstrating effectiveness with minimal data, as seen in systems developed with as little as one hour of conversational speech [8].

Despite these advancements, challenges remain in achieving expressive control over TTS outputs. Traditional systems often fail to capture human emotions and intentions, leading to speech that lacks expressiveness. Recent innovations integrating style and prosody modeling have addressed these limitations, enabling more dynamic and contextually appropriate speech synthesis [9].

The evolution of TTS technologies also addresses the linguistic needs of diverse communities. Initiatives like AfricanVoices aim to democratize access to TTS for low-resource languages through

community-driven approaches [10]. Similarly, the rapid development of deep learning-based TTS systems for Indian languages highlights the potential to serve a wide range of linguistic and cultural contexts [11]. Additionally, advancements in TTS for low-resource languages like Arabic emphasize the importance of efficient data utilization [12].

Interest is also growing in synthesizing non-traditional speech forms, such as rapping and singing, which present unique challenges due to their rhythmic and melodic complexities [13]. These developments illustrate the ongoing evolution of TTS technologies, continually pushing the boundaries of speech synthesis.

1.3 Applications of TTS Systems

TTS systems have diverse applications across multiple industries, significantly enhancing the accessibility and functionality of voice-based technologies. In telecommunications, TTS facilitates automated customer service and interactive voice response (IVR) systems, providing a natural communication interface between users and automated systems [7]. This technology is crucial for delivering consistent and understandable information, thereby improving customer satisfaction and reducing operational costs.

In education, TTS technologies are vital for creating inclusive learning environments. By converting written content into spoken words, TTS aids students with visual impairments and reading difficulties, ensuring equal access to educational materials. This application is particularly beneficial in language learning, where TTS can model correct pronunciation and intonation, providing learners with a reliable auditory reference [7].

TTS systems also play a significant role in assistive technologies for individuals with speech impairments, enhancing their communication abilities and promoting independence in daily activities. The integration of TTS into personal devices and communication aids underscores its importance in improving the quality of life for users with disabilities [7].

Moreover, TTS systems are increasingly utilized in e-commerce, especially in voice-based applications. In contexts involving code-mixed languages like Hindi-English, TTS is essential for accurately synthesizing product names and descriptions, thereby enhancing user experience and engagement in multilingual markets [14].

Recent research highlights the development of benchmark tools for TTS evaluation, facilitating automatic assessments of TTS quality. This advancement is crucial for refining TTS systems to better meet the specific needs and expectations of diverse user groups [15].

1.4 Structure of the Survey

This survey is organized to provide a comprehensive exploration of TTS systems, starting with foundational concepts and advancing to complex topics. The structure is as follows: Section 1 introduces the significance of TTS in human-computer interaction, its technological evolution, and diverse applications. Section 2 delves into background and definitions, clarifying key terminologies and contrasting parametric with neural TTS approaches. Section 3 discusses various speech synthesis architectures, emphasizing the role of deep learning in advancing these systems. Section 4 examines techniques for waveform generation and the evolution of vocoders, crucial for enhancing speech naturalness. Section 5 addresses prosody modeling, highlighting its challenges and the impact of deep learning on prosody enhancement. Section 6 explores advancements in multilingual and real-time TTS, focusing on global applications and processing challenges. Section 7 investigates voice conversion and zero-shot TTS, emphasizing their roles in personalization. Section 8 reviews evaluation metrics for TTS systems, underscoring the need for comprehensive benchmarks aligned with current research trends [16]. Finally, Section 9 presents the development of end-to-end TTS systems, detailing their integration with natural language processing techniques. The survey concludes with a discussion of future directions and potential global applications of TTS technologies.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Key Concepts and Terminologies in TTS

Text-to-Speech (TTS) systems rely on fundamental concepts to convert text into speech, with vocoders playing a pivotal role in shaping the quality and naturalness of synthesized outputs [17]. Prosody, which includes rhythm, stress, and intonation, is crucial for expressiveness and intelligibility, yet modeling it accurately remains challenging. Traditional phoneme-based methods often fall short in capturing prosodic variations, prompting advancements that incorporate syntactic information to improve naturalness [18, 19].

Contextual awareness is critical for end-to-end TTS systems, requiring the extraction of context from text inputs to produce coherent speech. Models like Vec-Tok Speech utilize semantic tokens and speech vectors to capture language intricacies [20]. Non-autoregressive models, such as NAR-TTS, enhance efficiency and reduce latency in synthesizing high-quality speech [21].

Frameworks like MASS integrate emotional and speaker voice conversion modules, enriching TTS outputs with expressive features [22]. However, many systems still struggle with accurately reflecting prosodic and emotional nuances [23]. The lack of high-quality data, especially for African languages, poses a significant challenge, necessitating innovative data collection strategies that blend real and synthetic sources [10, 16].

The reliance on annotated textual data can limit effectiveness due to text signal homogeneity. TTS systems are increasingly integrating diverse data sources and multimodal information, with pretrained language models (PLMs) enhancing prosody and pause predictions [24].

2.2 Parametric vs. Neural TTS

The shift from parametric to neural TTS systems represents a major evolution in speech synthesis methodologies. Early parametric systems used structured, rule-based frameworks, resulting in outputs that often lacked naturalness and expressiveness due to their reliance on handcrafted features [25, 26]. Neural TTS systems, leveraging deep learning, have transformed the field by enabling direct text-to-speech mapping, significantly enhancing naturalness and expressiveness [27]. Despite their advantages, neural systems face challenges such as high computational demands and the need for extensive datasets [28]. Intermediate feature generation inefficiencies can limit performance, especially in real-time applications [29].

Neural TTS systems excel in high-quality speaker adaptation and multilingual modeling, using high-resource language data to improve low-resource language synthesis. However, generating natural-sounding speech without extensive training data remains challenging, particularly in zero-shot contexts. Recent advancements in non-autoregressive models, like flow-based distributions and Diffusion Transformers, aim to improve efficiency and quality [30]. The TITW benchmark addresses training challenges with noisy, real-world data [31]. Optimizing TTS front-end models and integrating diverse audio information are crucial for enhancing synthesis effectiveness [32].

Challenges in training multi-speaker models with low-quality samples from crowdsourced datasets, coupled with noise-induced synthesis quality degradation, complicate neural TTS development [10, 29]. The inefficiency of multi-step sampling in diffusion models further hinders real-time capabilities [28]. As TTS systems evolve, the convergence of parametric and neural approaches may unlock new pathways for achieving high-quality, versatile speech synthesis across diverse applications and languages.

2.3 Role of Vocoders in TTS

Vocoder technology is integral to TTS systems, significantly affecting the quality and naturalness of synthesized speech by converting acoustic features into waveform signals. Traditional vocoders in statistical parametric synthesis often relied on simplified models, limiting expressiveness and naturalness [27]. Recent advancements have led to neural vocoders that surpass traditional models by capturing complex temporal dependencies and acoustic nuances [27]. However, challenges such as over-smoothing of features persist, potentially diminishing detail and expressiveness [33].

Innovations such as glottal vocoding enhance naturalness by parameterizing speech into glottal excitation and vocal tract components [34]. This nuanced representation improves synthesis quality. Vocoders play a critical role in TTS system benchmarking, evaluated alongside acoustic models to assess overall performance, guiding future advancements [27].

Developing universal vocoders capable of handling diverse acoustic conditions and speaker variations remains a challenge. Innovative design and training approaches are essential to ensure high-quality speech across applications and languages. As vocoder technology progresses, it will enhance the naturalness and expressiveness of synthetic speech, addressing the complexities of modeling highly expressive voices. Recent innovations, including Generative Adversarial Networks and Variational Auto-Encoders, are improving audio quality and narrowing the perceived naturalness gap between synthetic and human recordings [6, 35].

In recent years, the field of speech synthesis has witnessed significant advancements driven by deep learning technologies. These innovations have not only improved the quality of synthesized speech but have also enhanced the efficiency of text-to-speech (TTS) systems. To illustrate these developments, Figure 2 presents a comprehensive overview of the hierarchical structure of contemporary speech synthesis architectures. This figure highlights key neural TTS architectures and emphasizes innovative advancements in efficiency, quality, and acoustic modeling. The diagram categorizes major architectural features, advanced functionalities, and innovations in neural TTS systems, thereby showcasing the profound impact of deep learning on speech synthesis technology. By examining these elements, we can better understand the trajectory of research and development in this dynamic area.

3 Speech Synthesis Architectures

3.1 Neural TTS Architectures

Neural Text-to-Speech (TTS) architectures have revolutionized speech synthesis, leveraging deep learning to produce highly natural and expressive speech. Tacotron 2, a seminal architecture, employs a sequence-to-sequence model that directly maps text to speech waveforms, bypassing intermediate representations and enhancing output quality [36]. Non-attentive Tacotron models advance this by incorporating hierarchical prosodic controls for nuanced style transplantation [37].

EfficientSpeech exemplifies real-time synthesis optimization for ARM CPUs, demonstrating high-quality TTS in resource-limited environments [38]. SpeechX marks a significant leap by using a neural codec language model to generate acoustic tokens from textual and acoustic prompts [31]. Models like E2 TTS illustrate the move towards efficient, zero-shot TTS, converting text into character sequences and employing audio infilling tasks for mel spectrogram generation [30]. FastLTS combines a non-autoregressive acoustic decoder with a GAN-based vocoder for efficient high-quality synthesis [39].

Personalized TTS has advanced with architectures allowing precise control over accent and intensity, enhancing expressiveness [40]. Fish-Speech utilizes large language models and a Dual-AR architecture for multilingual and multi-emotional synthesis [41]. Integrating language models to predict future words improves prosody, making speech generation more coherent and contextually relevant [32]. HAM-TTS refines token-based systems with a latent variable sequence, enhancing robustness and flexibility [42].

ArtSpeech increases naturalness by modeling articulatory features [9]. OverFlow combines neural HMMs with normalizing flows for probabilistic speech acoustic modeling, improving synthesis accuracy [43]. DFSMN uses memory blocks to model long-term dependencies, enhancing system robustness [44]. NHMM-TTS integrates HMMs into sequence-to-sequence architectures, offering a probabilistic framework for speech generation [45].

DrawSpeech uses a sketch-conditioned diffusion model to create speech from user-drawn prosody sketches, translating them into detailed pitch and energy contours [46]. CosyVoice, a scalable TTS model, employs supervised semantic tokens for improved voice generation, enabling multilingual and zero-shot capabilities [1].

These advancements underscore the profound impact of deep learning on speech synthesis, expanding TTS applications and promoting linguistic and cultural inclusivity. The Conversational Context-

Aware End-to-End TTS (CCAE-TTS) method exemplifies this trend by integrating conversation context and auxiliary features to enhance synthesized speech naturalness [47].

As illustrated in Figure 4, neural TTS architectures can be hierarchically categorized into foundational models, advanced techniques, and innovative applications, showcasing significant advancements and applications in the field. The figure presents two key examples: "Speaker Classification Using Deep Learning and Audio Features," which demonstrates deep learning's capability to classify speakers by analyzing audio features, and "A diagram illustrating a speech synthesis process," which outlines a typical pipeline starting with audio input processed by LPCNet and culminating in synthetic speech generation. These examples highlight the intricate processes defining modern neural TTS architectures [48, 26].

3.2 Innovations in Neural TTS

Recent innovations in neural TTS systems have significantly enhanced training efficiency and synthesis quality. NaturalSpeech employs a variational autoencoder (VAE) architecture with phoneme pre-training and a memory mechanism to improve speech generation quality by capturing complex acoustic patterns [49]. The multi-rate attention architecture ensures constant synthesis speed and low latency while maintaining high-quality audio output [50].

EfficientSpeech reduces parameters and computational requirements, making high-quality TTS accessible on edge devices [38]. WaveCycleGAN offers a waveform-level approach, generating natural-sounding speech directly from synthetic inputs without relying on acoustic feature modeling [51].

Integrating normalizing flows into the NHMM framework, as seen in OverFlow, enhances acoustic distribution representation, enabling exact maximum likelihood training [43]. FastPitchFormant employs a decomposed structure to separately model formant and excitation representations, improving audio quality and pitch control [52].

F5-TTS introduces Sway Sampling at inference, enhancing text-speech alignment while maintaining naturalness and speaker similarity [53]. NHMM-TTS integrates HMMs with neural TTS, improving alignment and speech quality through better temporal dependency modeling [45].

The neural source-filter-based waveform model achieves significant waveform generation speed, maintaining quality comparable to autoregressive models [29]. PNG-BERT processes phonemes and graphemes simultaneously, improving interrelationship learning and addressing alignment issues [54].

These innovations collectively signify substantial advancements in neural TTS technology, resulting in systems that are more efficient, adaptable, and capable of producing speech closely resembling natural human voice. Advances in acoustic modeling, multi-speaker and multi-style training approaches, and deep convolutional networks streamline training while enhancing output quality, paving the way for diverse applications across various industries [55, 7, 56, 27, 57].

4 Waveform Generation and Vocoders

4.1 Techniques for Waveform Generation

Waveform generation is pivotal in TTS systems, bridging textual inputs with audible outputs. Traditional parametric models, such as LPCNet, relied heavily on handcrafted features and modular architectures, which limited their naturalness and expressiveness [27, 26]. The advent of deep learning has dramatically improved waveform generation, allowing for more flexible, data-driven approaches. The Neural Source-Filter Waveform Model (NSF) exemplifies this shift by using a sine-based excitation signal transformed into waveforms via a filter module, enhancing expressiveness [29].

Wave-Tacotron further advances this by generating waveforms directly from text using normalizing flows to model intra-waveform dependencies [58]. Similarly, SpeechX employs a neural codec language model to convert neural codes into waveforms, showcasing the novel approaches in this domain [31]. Techniques like CLONE integrate phoneme-level prosody modeling, ensuring high-

quality output by embedding prosodic elements into waveform generation [59]. ArtSpeech enhances voice tone and prosody similarity through articulatory representations [9].

The multi-band MelGAN (MBMelGAN) enhances efficiency and quality by generating waveforms from mel-spectrograms, processing sub-band signals, and integrating them into a full-band output [60]. This aligns with perceptual evaluations using perceptual loss based on Mean Opinion Score (MOS) [61]. DFSMN supports coherent and contextually appropriate speech generation by capturing long-term dependencies through a feed-forward architecture with memory blocks [44].

The transition from traditional parametric models to advanced neural approaches marks significant progress in TTS systems. Innovations driven by deep learning promise natural, expressive, and efficient speech synthesis across diverse applications. IA-TTS focuses on producing intelligible audio for individuals with hearing loss [62], while DelightfulTTS2 enhances synthesis quality through a codec network and acoustic model joint training [63]. WAVECYCLEGAN employs cycle-consistent adversarial networks to convert synthetic waveforms into natural ones, exemplifying diverse modern waveform generation techniques promising high-quality synthesis with reduced computational demands [51].

As shown in Figure 5, waveform generation and vocoder methodologies are crucial for advancing TTS systems. The first example emphasizes the interplay between acoustic models and waveform generators in producing coherent audio outputs, underscoring the importance of phase recovery for natural-sounding waveforms. The second example illustrates a TTS system flowchart, detailing the iterative nature of refining audio outputs. Lastly, the comparison of voice similarity across generative models quantitatively assesses their performance in replicating human-like speech characteristics, offering insights into model effectiveness in achieving high similarity scores [64, 65, 66].

4.2 Innovations in Vocoding and Waveform Generation

Recent innovations in vocoding and waveform generation have notably enhanced the naturalness and quality of synthesized speech, addressing challenges in TTS systems. The integration of generative adversarial networks (GANs) into vocoder design, as seen in WAVECYCLEGAN, mitigates the over-smoothing effect, enhancing naturalness through adversarial training [51]. WaveGlow highlights the shift towards real-time processing capabilities, converting mel-spectrograms into high-quality audio [67]. This underscores the need for efficient vocoding solutions that maintain high-quality synthesis.

The GAN-based glottal vocoder surpasses traditional DNN methods by capturing the stochastic nature of speech waveforms, enhancing expressiveness and fidelity [34]. OverFlow employs normalizing flows to model speech distributions, improving synthesis quality and convergence speed [43]. The NSF model accelerates synthesis by allowing simultaneous waveform sampling, optimizing vocoding techniques for enhanced performance and naturalness [29].

The feature smoothing augmentation method addresses mismatches between acoustic models and vocoders, ensuring high-quality outputs under varying conditions [33]. This contributes to more robust and versatile TTS systems. The effectiveness of voice conversion models in generating high-quality synthetic data that retains content information while enhancing style transfer is crucial for advancing vocoding techniques [4]. Adversarial training creates robust TTS models capable of generalizing to unseen speaker voices, exemplifying innovative vocoding approaches [68].

These innovations represent significant strides in vocoding and waveform generation, paving the way for more efficient, adaptable, and high-quality speech synthesis systems. As TTS technology progresses, particularly with deep learning integration and models like StyleTagging-TTS, these developments are poised to enhance performance substantially. This evolution promises increasingly natural and expressive speech outputs, facilitating diverse applications—from interactive media to voice conversion—while improving style control in speech synthesis [6, 69, 7].

Figure 6 illustrates these recent innovations in vocoding and waveform generation, categorizing them into generative models, normalizing flows, and training enhancements, thereby showcasing advancements in synthesized speech quality and efficiency.

5 Prosody Modeling

To effectively understand the complexities and advancements in prosody modeling, it is essential to first acknowledge the significant challenges that researchers and developers encounter in this domain. The intricate nature of prosody, which encompasses variations in pitch, duration, and intensity, complicates the synthesis process and impacts the overall quality of Text-to-Speech (TTS) outputs. In the subsequent subsection, we will delve into the specific challenges faced in prosody modeling, exploring the various factors that hinder the creation of natural and expressive synthesized speech.

5.1 Challenges in Prosody Modeling

Prosody modeling in Text-to-Speech (TTS) systems presents several significant challenges that affect the naturalness and expressiveness of synthesized speech. A primary issue is the difficulty in disentangling style and content, which leads to inaccuracies in synthesized speech and complicates the modeling of prosodic features [70]. Accurate prediction of phrase breaks is another critical challenge, as it is essential for maintaining intelligibility and naturalness in speech [71].

The mismatch between training and inference phases in universal vocoders and TTS acoustic models further complicates prosody modeling, often resulting in lower quality outputs [33]. Moreover, synthetic data generated for TTS may not perfectly match the target speaker's recordings, potentially affecting the quality and expressiveness of the synthesized speech [72].

Achieving precise control over prosody is essential for conveying intended meanings and emotions, yet remains a challenging task in TTS systems [46]. Evaluating the contextual appropriateness of prosody is also problematic, as current benchmarks may not adequately assess how well TTS systems generate prosodic variations that align with contextual cues [73].

Obtaining appropriate style reference clips is another obstacle, as these clips may not capture all the nuanced stylistic elements desired in the output speech [74]. Furthermore, attention mechanisms in neural TTS systems often fail to enforce a consistent ordering of speech sounds, which can lead to artifacts such as skipping and stuttering [45].

Lastly, existing evaluation methods do not sufficiently measure the diversity of synthetic speech compared to real speech, highlighting a key challenge in accurately modeling speech variability [5]. Despite the advancements offered by models like PnG BERT, which improve naturalness and effectively model prosody through phoneme and grapheme integration, significant challenges remain in achieving high-quality prosody modeling in TTS systems [54].

5.2 Methodologies for Prosody Manipulation

Prosody manipulation in Text-to-Speech (TTS) systems is crucial for enhancing speech expressiveness and achieving natural-sounding outputs. Various methodologies have been developed to address this challenge, focusing on different aspects of prosody modeling and synthesis. The hierarchical prosody modeling framework proposed by [75] conditions phoneme-level prosody predictions on word-level prosody predictions, thereby improving the synthesis process by capturing more nuanced prosodic variations. This hierarchical approach ensures a more coherent prosodic flow across speech segments.

An innovative method introduced by [76] involves target text prediction (TTP), which predicts prosody directly from linguistic representations, enhancing the voice conversion process by eliminating speaker mismatch. This approach provides a more accurate mapping of prosodic features to linguistic inputs, resulting in more natural speech synthesis.

The use of variational autoencoders (VAEs) to synthesize multiple prosodic renditions of a sentence, as explored by [77], addresses the average prosody issue by sampling from the tails of the prior distribution. This technique allows for the generation of diverse prosodic renditions, enriching the expressiveness of synthesized speech.

Incorporating a speech emotion recognition model as a style descriptor, as proposed by [78], enables TTS systems to learn prosody styles implicitly during training. This method leverages deep style features to enhance the emotional expressiveness of synthesized speech, making it more engaging and lifelike.

The use of a speech BERT model to extract prosody embeddings from speech segments, as suggested by [79], improves the prosody of synthesized speech by providing a robust representation of prosodic features. These embeddings facilitate more precise prosody manipulation, enhancing the overall quality of TTS outputs.

Prosody control mechanisms, such as those introduced by [80], allow TTS systems to predict and modify prosody info vectors based on linguistic input. This capability enables more expressive speech synthesis by providing fine-grained control over prosodic elements.

The DC CoMix TTS method, as described by [70], utilizes discrete codes as input to the reference encoder, enhancing the prosody transfer process by reducing content leakage. This approach ensures that the intended prosodic style is accurately conveyed in the synthesized speech.

Phrase Break Prediction Models (PBPM), proposed by [71], predict the locations of phrase breaks in text to improve the intelligibility of synthesized speech. Accurate phrase break prediction is essential for maintaining the natural rhythm and flow of speech, contributing to more intelligible and expressive TTS outputs.

Finally, the use of prosody sketches as control signals, as proposed by [46], allows users to indicate desired prosody trends through simple drawings. This user-friendly approach provides an intuitive means of controlling prosody, enhancing the flexibility and expressiveness of TTS systems.

The methodologies discussed in these references collectively signify substantial progress in the manipulation of prosodic parameters within Text-to-Speech (TTS) systems. By employing advanced speech processing techniques, such as feature extraction and synthesis, researchers have developed frameworks that enhance the naturalness and expressiveness of TTS outputs. These advancements include the integration of emotional information into speaker embeddings, the establishment of a unified front-end framework for improved linguistic feature extraction, and novel evaluation methods that allow for more precise assessments of prosodic variations. Consequently, these innovations contribute to the ongoing refinement of TTS systems, enabling them to produce speech that more closely resembles the nuances of human communication. [7, 81, 23, 82, 73]

As shown in Figure 7, this figure illustrates key methodologies for prosody manipulation in TTS systems, categorized into hierarchical approaches, generative models, and prosody embedding techniques. Each category showcases different methods aimed at enhancing speech expressiveness and naturalness. In the realm of speech synthesis, prosody modeling plays a pivotal role in enhancing the naturalness and expressiveness of generated speech. Prosody, which encompasses elements such as intonation, rhythm, and stress, is essential for conveying meaning and emotion in spoken language. The methodologies for prosody manipulation are diverse, as illustrated by the examples provided in Figure **??**. The first example focuses on a mel-spectrogram-based speech synthesis system, which utilizes a decoder and WaveRNN to transform mel-spectrograms into speech while employing various predictors and controllers for prosody manipulation. The second example delves into the training and conversion processes of speech synthesis systems, highlighting the stages of recognizer training, synthesizer pretraining, and fine-tuning with multi-speaker datasets to achieve accurate text-to-speech conversion. Lastly, a flowchart exemplifies a speech synthesis process that begins with phonemes, which are processed by a text encoder and prosody predictor to determine the duration and pitch, ultimately resulting in refined speech output. Together, these methodologies underscore the complexity and importance of prosody in modern speech synthesis systems. [83, 76, 23]

5.3 Role of Deep Learning in Prosody Enhancement

Deep learning techniques have significantly advanced prosody modeling in Text-to-Speech (TTS) systems, offering sophisticated tools to enhance the naturalness and expressiveness of synthesized speech. A noteworthy innovation is the use of prosody encoders, which extract prosody information separately from text, allowing for more natural speech synthesis compared to traditional methods that rely solely on linguistic features [84]. This separation enables the modeling of intricate prosodic variations, resulting in more expressive speech outputs.

The MQTTS framework exemplifies the power of deep learning by leveraging multiple codebooks to capture diverse prosody patterns, thereby improving the robustness of synthesis against noise [3]. This approach ensures that synthesized speech maintains its quality even in challenging acoustic conditions, enhancing the listener's experience.

Advanced neural codec architectures, as proposed by [85], further enhance the model's understanding of speech features, improving prosody and expressiveness. These architectures facilitate the capture of complex prosodic patterns, leading to more natural and engaging speech synthesis.

The integration of multi-task adversarial training, as described by [68], improves the quality of speech synthesis for both seen and unseen speakers. This method enhances the model's adaptability, allowing it to produce high-quality prosody across various speaker profiles.

Moreover, the use of conversational context encoders, as highlighted by [47], enhances prosody and context-awareness by utilizing auxiliary encoders to capture conversational nuances. This approach ensures that synthesized speech is contextually appropriate and resonates well with human listeners.

The ability to leverage the richness of natural language to guide stylistic transformations in voice conversion, as demonstrated by [74], further underscores the potential of deep learning in achieving expressive speech synthesis. This capability is crucial for producing speech that reflects a wide range of stylistic and emotional nuances.

These advancements collectively highlight the transformative impact of deep learning on prosody enhancement, paving the way for more natural, expressive, and contextually appropriate TTS systems. As research advances, the integration of deep learning techniques into prosody modeling is set to significantly enhance the quality and effectiveness of speech synthesis technologies. This evolution is marked by the transition from traditional deep neural network-based methods to advanced sequence-to-sequence models with attention mechanisms, which have achieved near-human speech quality. Notably, the incorporation of phonetic and prosodic features, particularly for pitch-accented languages like Japanese, is crucial for improving pronunciation accuracy. Furthermore, recent developments in automatic prosody control allow for fine-tuning of speech expressiveness and pacing, thereby facilitating a more natural and adaptable speech synthesis experience. [80, 86]

6 Multilingual and Real-Time TTS

6.1 Advancements in Multilingual TTS Systems

Recent developments in multilingual Text-to-Speech (TTS) systems have significantly enhanced the synthesis of speech across various languages, addressing linguistic diversity challenges. The Flow-Matching-Based Noise-Robust Zero-Shot TTS (FM-NR-ZSTTS) exemplifies this progress by using flow-matching techniques to produce clean speech from noisy audio prompts, demonstrating robustness in multilingual settings [87]. Similarly, the CosyVoice model showcases scalability by generating high-quality speech with improved semantic alignment and speaker characteristics, highlighting its potential for diverse linguistic applications [1].

These advancements are furthered by methodologies that leverage cross-lingual transferability of multilingual language models, facilitating TTS development for low-resource languages using text-only data [88]. The YourTTS framework enhances zero-shot multi-speaker TTS and voice conversion through a multilingual approach, allowing effective synthesis with minimal data [89]. This is crucial for high-quality voice conversion, as deep learning techniques significantly improve conversion quality [90].

Additionally, the multi-task adversarial training algorithm enhances synthetic speech quality for unseen speakers, proving effective in multi-speaker TTS applications [68]. The integration of large multilingual datasets for pretraining exemplifies further advancements, facilitating natural and expressive speech synthesis across multiple languages [85]. Together, these developments mark significant progress in creating more natural, expressive, and accessible speech synthesis, promising high-quality outputs across diverse linguistic landscapes.

6.2 Challenges in Real-Time Processing

Real-time Text-to-Speech (TTS) synthesis faces numerous challenges due to the high computational demands and inefficiencies of existing models. Latency is a major issue, especially with CPU-based devices, where TTS systems struggle to synthesize speech in real-time [27]. Zero-shot TTS systems exacerbate these challenges, often failing to synthesize speech promptly from text streams [91]. The entangled nature of speech components complicates the learning of generalized features needed for effective synthesis [92].

As illustrated in Figure 8, the challenges faced by real-time TTS systems can be categorized into latency issues, scalability problems, and innovative solutions. Each category highlights specific aspects such as CPU-based latency, adaptation to new speakers, and advanced architectures like multi-rate attention and ParaNet for improved performance. Adapting TTS systems to new speakers or languages without sufficient parallel data limits scalability, particularly for low-resource languages with scarce audio recordings [93, 88]. The MoA-TTS method addresses these challenges by operating on edge devices, yet existing models often degrade in voice quality when style prompt distributions differ from training data [94, 95]. The main obstacle remains the insufficient quality of current zero-shot TTS models, which struggle to generalize to speakers with different voice characteristics [89].

Innovative architectures, such as the multi-rate attention architecture, maintain constant latency and real-time factor (RTF) regardless of input length, ensuring consistent performance in real-time applications [50]. ParaNet achieves real-time synthesis, significantly reducing latency by operating 254.6 times faster than real-time [28]. Despite these advancements, methods like HAM-TTS still require optimization for practical usability [42]. SpeechX addresses real-time processing challenges by utilizing a unified model for diverse audio-text speech generation tasks, including zero-shot TTS and noise suppression [31].

Future research may explore pruning techniques and additional datasets to enhance real-time processing capabilities in multilingual or multi-speaker TTS systems [36]. Leveraging these advancements can help TTS systems achieve the responsiveness and quality necessary for real-time applications across diverse linguistic and technological contexts.

6.3 Techniques for Efficient TTS Synthesis

Efficient Text-to-Speech (TTS) synthesis is crucial for optimizing systems to achieve faster processing and reduced latency without compromising speech quality. Recent advancements have focused on models and methodologies suitable for real-time applications. FLY-TTS, for instance, achieves a real-time factor of 0.0139 on CPU, surpassing existing models and proving its suitability for real-time scenarios [96]. FastPitchFormant offers breakthroughs in audio quality and pitch control, especially with large pitch shifts, outperforming traditional methods [52].

Innovative approaches like streaming inference have significantly reduced latency by enabling real-time processing of continuous text streams. End-to-end TTS models with autoregressive attention-based architectures and LPCNet vocoders achieve nearly constant latency, enabling efficient feature generation at speeds up to 31 times faster than real-time on CPUs. The Interleaved Speech-Text Language Model (IST-LM) enhances streaming capabilities by training directly on interleaved text and speech sequences, streamlining synthesis without complex duration predictions. Similarly, the multi-rate attention architecture maintains low latency and real-time factor (RTF) regardless of input length, making these systems well-suited for immediate audio responses [97, 98, 50, 99].

The multi-rate attention model optimizes TTS systems by employing dynamic pooling to limit attention context length, enhancing processing speed without sacrificing speech quality. This method effectively addresses latency and real-time factor (RTF) challenges associated with traditional high-quality TTS systems, ensuring a balance between synthesis speed and audio quality [7, 50].

Recent advancements also incorporate phoneme prediction and monotonic alignment strategies, significantly improving robustness and efficiency. The alignment mechanism from RAD-TTS enhances phoneme alignments, making TTS systems more resilient to errors in long utterances and out-of-domain text. Additionally, the ESPnet2-TTS toolkit facilitates joint training with neural vocoders, offering state-of-the-art models that streamline training processes, resulting in higher synthesis quality that rivals ground-truth speech outputs [100, 101].

The methodologies discussed reflect substantial advancements in TTS technology, particularly through approaches like IST-LM and synthesis techniques, including neural and hybrid TTS. These advancements enhance the naturalness and intelligibility of synthesized speech while significantly improving processing speed and reducing latency, enabling efficient real-time applications in human-computer communication [97, 7]. As research evolves, these techniques promise increasingly efficient and versatile TTS systems capable of meeting real-time application demands across diverse platforms and languages.

7 Voice Conversion and Zero-Shot TTS

7.1 Techniques for Voice Conversion

Voice conversion (VC) in Text-to-Speech (TTS) systems involves modifying a speaker's voice to mimic another's, utilizing varied methodologies for natural transformation. Traditional VC approaches include parallel techniques requiring matched source-target speech pairs and non-parallel methods that use unpaired data, enhancing flexibility [102]. A notable advancement is the Comprehensive Voice Conversion Framework (CVCF), which uses deep neural networks to convert high-resolution spectral and prosodic features such as fundamental frequency (F0), intensity, and duration [103].

The Bootstrapping Voice Conversion from Speaker-Adaptive TTS (BVC-TTS) method demonstrates effective performance with minimal target speaker data, broadening applicability across diverse linguistic contexts [93]. This approach underscores adaptive techniques that enhance VC flexibility and scalability.

Recent deep learning advancements have introduced style-based TTS models that learn disentangled speech representations through transfer learning, enabling one-shot voice conversion without input text [104]. These frameworks facilitate rapid adaptation to new speakers, increasing VC application versatility.

Integrating automatic speech recognition (ASR) with Transformer TTS models, enhanced by prosody encoders, improves speech naturalness by capturing intricate prosodic variations [84]. This cascading approach underscores prosody's importance in achieving high-quality voice conversion.

Innovative architectures, such as fully convolutional models, enable zero-shot voice conversion, allowing transformation between speakers without parallel datasets [67]. This capability is essential for applications requiring quick adaptation to new speakers with minimal data.

The survey by [105] categorizes voice conversion methods into three stages: speech analysis, mapping, and reconstruction/synthesis, highlighting the role of deep learning architectures like autoencoders (AEs), variational autoencoders (VAEs), and generative adversarial networks (GANs) in advancing VC technologies.

Furthermore, the proposed voice conversion model by [74] enhances style transformation flexibility by processing a wide range of textual instructions, allowing for nuanced and personalized voice conversion outcomes. These diverse techniques collectively represent ongoing advancements in voice conversion, paving the way for more natural, adaptable, and efficient TTS systems capable of transforming voices across various linguistic and emotional contexts.

7.2 Integration of Voice Conversion and Zero-Shot TTS

Integrating voice conversion techniques with zero-shot Text-to-Speech (TTS) systems marks a pivotal advancement in speech synthesis, enabling speech generation without prior exposure to a speaker's voice. Models like ConVoice exemplify effective zero-shot voice conversion capabilities, functioning without parallel or transcribed data [67]. Such models enhance zero-shot TTS systems' adaptability and scalability in multi-speaker scenarios with minimal data.

A notable development in this domain is the use of transfer learning frameworks that incorporate robust linguistic representations from TTS training into voice conversion models, improving synthesis quality in zero-shot TTS applications. The UnifySpeech framework further enhances zero-shot TTS capabilities by allowing simultaneous speech generation and transformation through VC within a unified model. This model decouples speech into three independent components—content, speaker, and prosody—enabling sophisticated, context-aware speech synthesis without requiring target speaker training data. Consequently, the system improves speaker modeling and content decoupling, leading to versatile speech synthesis applications [106, 107].

Innovative approaches like the Interleaved Speech-Text Language Model (IST-LM) significantly enhance the integration of voice conversion within zero-shot TTS systems. The IST-LM is trained on interleaved sequences of text and speech tokens at a fixed ratio, streamlining processes by eliminating the need for complex tasks such as duration prediction and grapheme-to-phoneme alignment. Its effectiveness relies on factors like the spatial relationship between speech and text tokens and the

accessibility of subsequent text tokens for each speech token. Experimental results indicate that IST-LM achieves optimal streaming TTS performance with minimal engineering overhead, enhancing real-time text streaming from large language models while preserving naturalness and intelligibility in generated speech [55, 97, 108, 88, 48]. This method ensures synthesized speech remains coherent and contextually appropriate, maintaining the original speaker's voice expressiveness.

Despite these advancements, challenges persist in aligning non-parallel data, crucial for maintaining high-quality, identity-preserving speech synthesis across diverse linguistic and technological contexts. Addressing these challenges necessitates innovative solutions that enhance the flexibility and scalability of integrated voice conversion and zero-shot TTS systems. As deep learning and artificial intelligence continue to advance the field, integrated TTS systems are positioned to deliver high-quality, human-like speech synthesis for various applications and user requirements. These systems leverage sophisticated components, including text analysis, acoustic modeling, and vocoding, increasingly capable of producing expressive and adaptive speech. Recent innovations, such as Tacotron 2, WaveNet, and FastSpeech, exemplify this progress, offering substantial improvements in accuracy and accessibility for diverse media applications, from dubbing to narration [6, 81, 27].

8 Evaluation Metrics for TTS

Evaluating Text-to-Speech (TTS) systems requires a nuanced understanding of diverse metrics that measure performance through both objective and subjective lenses. This section delves into these evaluation metrics, highlighting their roles in assessing the accuracy, naturalness, and overall quality of synthesized speech. The following subsection will focus on objective evaluation metrics, which are crucial for quantitatively benchmarking TTS systems against established standards.

8.1 Objective Evaluation Metrics

Benchmark	Size	Domain	Task Format	Metric
TTS-Bench[64]	50,000	Speech Synthesis	Speech Quality Evaluation	MOS, F0
ML-TTS[109]	1,000,000	Text-to-Speech Synthesis	Voice Synthesis	Mean Opinion Score
TITW[110]	530,630	Text-To-Speech Synthesis	Speech Synthesis	MCD, DNSMOS
ESPnet2-TTS[101]	24,000	Text-to-Speech	Speech Synthesis	MCD, F0 RMSE
MLMS[111]	80,000	Neural Speech Synthesis	Speech Naturalness Evaluation	MUSHRA
TTS-QA[15]	9,600	Text-to-Speech Synthesis	Quality Assessment	SRCC, KTAU
TTS-Benchmark[112]	13,100	Text-to-Speech	Phoneme Classification	Accuracy
FAinASR[113]	21,925	Speech Recognition	False Alarm Detection	Word Error Rate, F1 Score

Table 1: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of text-to-speech (TTS) and speech synthesis systems. It details the benchmark names, dataset sizes, domains, task formats, and the metrics employed for assessment, providing a valuable resource for comparing different TTS methodologies and their performance metrics.

Objective evaluation metrics are essential for quantitatively assessing TTS performance, providing a standardized framework for model comparison. Character Error Rate (CER) and Word Error Rate (WER) are pivotal in evaluating the accuracy of TTS outputs processed by Automatic Speech Recognition (ASR) systems, reflecting the system's ability to reproduce nuanced human speech [12]. Metrics such as Mean Square Error (MSE), Linear Correlation Coefficient (LCC), Spearman's Rank Correlation Coefficient (SRCC), and Kendall Tau Rank Correlation (KTAU) measure the correlation between true and predicted Mean Opinion Scores (MOS), offering insights into the perceived quality of synthesized speech [114]. These statistical measures are vital for aligning objective outputs with human perceptual evaluations. Table 1 provides a detailed summary of key benchmarks used in the objective evaluation of TTS systems, highlighting the diversity in dataset sizes, application domains, task formats, and evaluation metrics.

Subjective assessments, particularly MOS and Comparative Mean Opinion Scores (CMOS), derive from trained evaluators who assess audio quality, reflecting human judgment on naturalness and intelligibility. Recent advancements have introduced automated methods utilizing neural networks trained on human MOS ratings to predict audio quality efficiently, revealing variability in speaker contributions to perceived quality and aiding in identifying high-quality speakers for TTS system development [61, 15]. The integration of subjective and objective metrics fosters a comprehensive

evaluation of TTS systems, balancing technical performance with user satisfaction. CMOS tests, where native speakers rate the prosodic performance of synthesized speech, provide additional quantitative measures for evaluation [47]. By amalgamating these diverse metrics, researchers can effectively assess TTS system performance, facilitating the development of high-quality speech synthesis technologies tailored to diverse user needs across various applications and linguistic contexts.

8.2 Subjective Evaluation Methods

Subjective evaluation methods are crucial for assessing the perceptual quality of TTS outputs, offering insights into listener perceptions of naturalness, expressiveness, and overall quality. The Mean Opinion Score (MOS) test is a primary approach, where listeners rate speech quality on a scale from 1 to 5, with higher scores indicating greater naturalness and satisfaction. MOS assessments are often complemented by Best Worst Scaling (BWS), which provides comparative quality measures by asking participants to select the best and worst samples from a set [40].

Subjective evaluations may also involve detailed listening tests where participants assess specific attributes of synthesized speech, such as speaker similarity, prosody, and intelligibility. Effectiveness can be gauged through metrics like speaker identification accuracy and the intelligibility of reconstructed waveforms, providing a comprehensive understanding of system performance [115]. These tests are essential for identifying strengths and weaknesses in TTS systems, guiding further improvements in synthesis quality.

Moreover, subjective evaluations often incorporate linguistic features, such as part-of-speech (POS) tags and BERT embeddings, to enhance synthetic speech assessment. By augmenting existing MOS prediction models with prosodic features like phoneme-level F0 and duration, researchers can better understand how these elements contribute to perceived speech quality [114].

The integration of subjective and objective evaluation methods, including Speaker Encoder Cosine Similarity (SECS) and Mel-Cepstral Distortion (MCD), provides a holistic approach to TTS assessment. This combination enables a thorough evaluation of both technical accuracy and perceptual quality, with correlations between subjective scores and benchmarks like TTSDS validating these methods' effectiveness [116]. Subjective evaluation methods are crucial for developing and enhancing TTS systems, ensuring alignment with listeners' perceptual expectations and producing high-quality, natural-sounding speech. Approaches like MOS and innovative paradigms such as Rapid Prosody Transcription not only assess overall quality but also identify specific synthesis errors, particularly in prosodic variations. Recent advancements in TTS technologies, including neural and hybrid models, have underscored the importance of rigorous subjective evaluations to achieve human-level quality and effectively adapt TTS systems for diverse applications [49, 7, 73].

8.3 Hybrid Evaluation Approaches

Hybrid evaluation approaches for TTS systems integrate objective and subjective metrics, providing a comprehensive framework for assessing synthesized speech quality and effectiveness. These approaches leverage the strengths of each evaluation type, offering nuanced insights into TTS performance. Objective metrics, such as CER, WER, and MCD, yield quantitative measures of accuracy and fidelity, essential for benchmarking TTS systems against established standards [73].

Subjective evaluations, including MOS and BWS, capture listener perceptions of synthetic speech, assessing key attributes like naturalness and expressiveness. However, MOS tests primarily provide a general assessment of overall quality, lacking the granularity needed to identify specific synthesis errors, particularly in prosodic variation. To enhance evaluation accuracy, methods like Rapid Prosody Transcription have been proposed, enabling listeners to pinpoint errors in real-time and offering a more detailed representation of perceptual inaccuracies in speech synthesis. This approach correlates with traditional MOS rankings while revealing nuanced insights into the effectiveness of prosodic features in various contexts, such as neural TTS systems generating contextually appropriate prosodic prominence [61, 15, 73]. This duality allows researchers to identify discrepancies between technical performance and user satisfaction, guiding refinements in TTS models.

Recent advancements have introduced innovative benchmarks that enhance traditional evaluation methods. For example, a new benchmark incorporates a real-time error marking task, enabling

listeners to identify specific prosodic errors, thus providing a more detailed assessment of prosody, often inadequately captured by conventional metrics [73]. By allowing listeners to annotate specific errors, this benchmark offers valuable insights into the prosodic quality of TTS outputs, facilitating targeted improvements in synthesis quality.

The integration of hybrid evaluation approaches ensures a balanced assessment of TTS systems, combining the precision of objective metrics with the perceptual insights of subjective evaluations. This comprehensive evaluation framework is essential for advancing TTS technologies, ensuring compliance with rigorous technical standards while prioritizing user experience across various applications and linguistic contexts. It incorporates critical factors such as voice naturalness, prosody, speaker identity, and intelligibility, informed by recent research highlighting the interdependencies among TTS components, including text normalization, prosody modeling, and acoustic synthesis. By integrating these elements, the framework aims to enhance the overall performance and applicability of TTS technologies in modern interactive media and speech-to-speech translation tasks [7, 108, 6, 116, 81].

9 End-to-End TTS Systems

9.1 Advancements in End-to-End TTS Architectures

Recent progress in end-to-end Text-to-Speech (TTS) architectures has significantly enhanced both the efficiency and quality of speech synthesis, surpassing traditional modular frameworks. A notable development is the shared decoder architecture, which simultaneously processes multiple input sources to improve performance while reducing computational redundancy [117]. This innovation streamlines synthesis, producing more coherent and natural speech outputs.

Models such as Wave-Tacotron demonstrate the benefits of direct waveform generation from text, eliminating intermediate features and enhancing generation speed through parallelization techniques like normalizing flows [58]. FLY-TTS, employing ConvNeXt blocks and grouped parameter-sharing, optimizes inference speed and minimizes model size without sacrificing speech quality, proving the feasibility of deploying high-performance TTS systems in resource-constrained environments [96]. The integration of advanced vocoders like MBMelGAN has further improved speech generation quality and efficiency, achieving a real-time factor of 0.03 [60].

The Tacotron-PL training strategy has shown substantial improvements over existing Tacotron and GST-Tacotron baselines, particularly in expressiveness and naturalness [78]. This strategy highlights broader applications for end-to-end neural TTS systems, facilitating the synthesis of more engaging and lifelike speech. Additionally, incorporating a trained phrasing model enhances listener comprehension, especially in complex narratives [71], underscoring the importance of contextual and prosodic modeling in achieving high-quality speech synthesis.

Future research may focus on integrating advanced vocoding techniques, such as the ExcitNet vocoder, to further enhance modeling capabilities and synthesis quality [118]. Methods like the Modified Differential Multiplier Method (MDMM) offer promising avenues for optimizing end-to-end TTS architectures by constraining reconstruction loss without hyper-parameter tuning [119].

Innovations such as the Interleaved Speech-Text Language Model (IST-LM) and Easy End-to-End Diffusion-based TTS (E3 TTS) showcase the potential of integrated models to transform speech synthesis. IST-LM simplifies the TTS process by training on interleaved text and speech tokens, reducing complexities related to duration prediction while maintaining performance akin to traditional non-streaming systems. In contrast, E3 TTS employs a diffusion-based approach to generate audio waveforms directly from text, enhancing flexibility in zero-shot tasks. Collectively, these advancements demonstrate the capability of modern TTS systems to efficiently produce high-quality, natural-sounding speech across diverse applications and linguistic contexts [97, 120].

9.2 Integration of Natural Language Processing Techniques

The integration of Natural Language Processing (NLP) techniques into end-to-end Text-to-Speech (TTS) systems has significantly enriched linguistic processing capabilities, enabling more natural and contextually appropriate speech synthesis. The Tacotron-PL training strategy optimizes frame-level spectral loss and utterance-level style loss in Tacotron-based systems, enhancing synthesized speech

expressiveness [78]. This dual optimization allows TTS systems to effectively capture and reproduce the stylistic nuances of human speech.

Incorporating NLP techniques such as semantic parsing and syntactic analysis facilitates accurate textual interpretation, allowing for precise prosody and intonation modeling. Advanced NLP algorithms enable TTS systems to interpret contextual and emotional nuances, generating speech that is both expressive and aligned with natural human prosody. Techniques like prosodic parameter manipulation, emotion-aware speaker embeddings, and style tagging enhance the system's ability to produce audio that mimics the variability and richness of human expression, thereby improving user engagement [6, 112, 69, 23, 82].

The application of pretrained language models (PLMs) in TTS systems has further improved linguistic processing, enhancing prosody and yielding more natural-sounding speech. Studies indicate that PLMs can effectively predict prosody and pauses, demonstrating a logarithmic relationship between model size and output quality, alongside performance variations between neutral and expressive speech. The development of unified front-end frameworks for TTS underscores the interdependencies among linguistic feature extraction modules, leading to state-of-the-art results in speech synthesis. PLMs provide rich contextual information, enabling TTS systems to generate syntactically correct and semantically meaningful speech, effectively handling complex linguistic structures and idiomatic expressions [55, 121, 24, 81].

Additionally, NLP techniques such as sentiment analysis and emotion detection are increasingly utilized in TTS systems to tailor speech synthesis to the emotional tone of input text. By integrating advanced emotional and prosodic parameter manipulation techniques, TTS systems can produce speech that reflects the intended emotional context while closely mimicking natural human prosody. This enhancement leads to more expressive outputs, increasing listener engagement and improving synthesized speech quality. Research has shown that emotional speaker embeddings and style tagging significantly elevate TTS systems' expressiveness, making them more intuitive and effective in conveying nuanced emotional tones [82, 69, 23].

The integration of advanced NLP techniques into end-to-end TTS systems represents a significant advancement in speech synthesis technology. This innovation facilitates the automatic extraction and utilization of linguistic and contextual features—such as vowel reduction, lexical stress, and part-of-speech tagging—resulting in high-quality, natural-sounding, and contextually relevant speech. These enhancements enable TTS systems to operate effectively across various applications and linguistic environments, achieving human-level quality in synthesized speech, as evidenced by recent benchmarks [49, 27, 6, 81, 112]. The ongoing synergy between NLP and TTS technologies promises increasingly sophisticated and versatile speech synthesis solutions.

10 Conclusion

10.1 Future Directions and Global Applications

The trajectory of Text-to-Speech (TTS) systems is poised for significant innovation, emphasizing enhanced expressiveness, adaptability, and efficiency across varied applications. A primary focus lies in developing data-efficient models capable of generating expressive speech with minimal resources, crucial for low-resource and underrepresented languages. Advancements in TTS architectures, such as the integration of Transformer models and multi-speaker synthesis, are pivotal for technological evolution. Enhancing pronunciation styles and mitigating background noise through adaptive techniques are also essential for improving performance in challenging environments.

Future research endeavors should refine training methodologies and integrate additional prosodic features to improve model generalization and phrase break prediction. The optimization of Generative Adversarial Network (GAN) architectures for real-time applications presents a promising path forward, enabling more natural and coherent speech synthesis. Expanding voice conversion techniques to encompass diverse languages and styles will further broaden TTS applications. Acquiring more data and leveraging self-supervised representations will enhance synthesis quality, producing nuanced and expressive outputs.

Further, optimizing model architectures, diversifying training datasets, and enhancing pitch control robustness are critical for advancing TTS technologies. Emphasis should also be placed on improving alignment accuracy and integrating additional features to enhance model capabilities. Exploring

multi-speaker functionalities and incorporating style features in systems like FLY-TTS will diversify speech synthesis options.

Research aimed at improving the interpretability of voice conversion systems and developing efficient real-time models is vital for expanding TTS applicability. Investigating robust network designs, including transformers and probabilistic post-nets, can enhance the naturalness of synthesized speech. Incorporating traditional speech modeling methods and simplified convolution blocks may further bolster model performance.

Future directions also include refining speaker-adaptive methods and exploring their integration with end-to-end TTS systems. Expanding model capabilities to additional languages and leveraging them for data augmentation in automatic speech recognition is crucial, especially in low-resource contexts. Optimizing data generation and model training for improved performance and stability, alongside exploring applications in other low-resource languages, remains a promising research avenue. Enhancing zero-shot scenario performance and exploring alternative vocoding methods to improve audio quality are key areas of focus. Additionally, improving adaptability to diverse accents in code-mixed speech and advancing pre-training methodologies will propel the field forward.

Exploring cross-language expressive datasets for low-resource environments through synthetic augmentation presents another promising direction. Automating data selection processes and applying methods to other low-resource languages are areas ripe for development. Experimentation with model architectures and encoding strategies, alongside practical applications of MOS prediction models for neural TTS evaluation, are crucial next steps. Expanding corpus size and refining models to enhance conversational expressiveness are also potential research areas.

These research directions, coupled with efforts to enhance global accessibility and explore new applications, highlight the transformative potential of TTS technologies in bridging linguistic and cultural divides, promoting inclusivity, and facilitating more natural human-computer interactions worldwide. Improving synthesized speech quality and diversifying instruction types in voice conversion systems will further the evolution of TTS technologies.

References

[1] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024.

[2] Raviraj Joshi and Nikesh Garera. Rapid speaker adaptation in low resource text to speech systems using synthetic data and transfer learning, 2023.

[3] Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky. A vector quantized approach for text to speech synthesis on real-world spontaneous speech, 2023.

[4] Lucas H. Ueda, Leonardo B. de M. M. Marques, Flávio O. Simões, Mário U. Neto, Fernando Runstein, Bianca Dal Bó, and Paula D. P. Costa. Exploring synthetic data for cross-speaker style transfer in style representation based tts, 2024.

[5] Christoph Minixhofer, Ondřej Klejch, and Peter Bell. Evaluating and reducing the distance between synthetic and real speech distributions, 2023.

[6] Mohammad Reza Hasanabadi. An overview of text-to-speech systems and media applications, 2023.

[7] Md. Jalal Uddin Chowdhury and Ashab Hussan. A review-based study on different text-to-speech technologies, 2023.

[8] Giulia Comini, Goeric Huybrechts, Manuel Sam Ribeiro, Adam Gabrys, and Jaime Lorenzo-Trueba. Low-data? no problem: low-resource, language-agnostic conversational text-to-speech via f0-conditioned data augmentation, 2022.

[9] Zhongxu Wang, Yujia Wang, Mingzhu Li, and Hua Huang. Artspeech: Adaptive text-to-speech synthesis with articulatory representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 535–544, 2024.

[10] Perez Ogayo, Graham Neubig, and Alan W Black. Building african voices, 2022.

[11] Gokul Karthik Kumar, SV Praveen, Pratyush Kumar, Mitesh M Khapra, and Karthik Nandakumar. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[12] Massa Baali, Tomoki Hayashi, Hamdy Mubarak, Soumi Maiti, Shinji Watanabe, Wassim El-Hajj, and Ahmed Ali. Unsupervised data selection for tts: Using arabic broadcast news as a case study, 2023.

[13] Konstantinos Markopoulos, Nikolaos Ellinas, Alexandra Vioni, Myrsini Christidou, Panos Kakoulidis, Georgios Vamvoukakis, Georgia Maniati, June Sig Sung, Hyoungmin Park, Pirros Tsiakoulis, and Aimilios Chalamandaris. Rapping-singing voice synthesis based on phoneme-level prosody control, 2021.

[14] Raviraj Joshi and Nikesh Garera. Code-mixed text to speech synthesis under low-resource constraints, 2023.

[15] Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King. Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis, 2020.

[16] Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Klimkov Viacheslav. Effect of data reduction on sequence-to-sequence neural tts, 2018.

[17] Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*, 2024.

[18] Zhen Zeng, Jianzong Wang, Ning Cheng, and Jing Xiao. Prosody learning mechanism for speech synthesis system without text length limit, 2020.

[19] Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu. Syntaspeech: Syntax-aware generative adversarial text-to-speech, 2022.

[20] Xinfa Zhu, Yuanjun Lv, Yi Lei, Tao Li, Wendi He, Hongbin Zhou, Heng Lu, and Lei Xie. Vec-tok speech: speech vectorization and tokenization for neural speech generation. *arXiv preprint arXiv:2310.07246*, 2023.

[21] Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech, 2022.

[22] Jinyin Chen, Linhui Ye, and Zhaoyan Ming. Mass: Multi-task anthropomorphic speech synthesis framework, 2021.

[23] Exploiting emotion information i.

[24] Marcel Granero-Moya, Penny Karanasou, Sri Karlapati, Bastian Schnell, Nicole Peinelt, Alexis Moinet, and Thomas Drugman. A comparative analysis of pretrained language models for text-to-speech, 2023.

[25] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks, 2017.

[26] Zvi Kons, Slava Shechtman, Alex Sorin, Carmel Rabinovitz, and Ron Hoory. High quality, lightweight and adaptable tts using lpcnet, 2019.

[27] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis, 2021.

[28] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. Non-autoregressive neural text-to-speech, 2020.

[29] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis, 2019.

[30] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts, 2024.

[31] Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. Speechx: Neural codec language model as a versatile speech transformer, 2024.

[32] Brooke Stephenson, Thomas Hueber, Laurent Girin, and Laurent Besacier. Alternate endings: Improving prosody for incremental neural tts with predicted future text input, 2021.

[33] Jeongmin Liu and Eunwoo Song. Training universal vocoders with feature smoothing-based augmentation methods for high-quality tts systems, 2024.

[34] Bajibabu Bollepalli, Lauri Juvela, and Paavo Alku. Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis, 2019.

[35] Abdelhamid Ezzerg, Adam Gabrys, Bartosz Putrycz, Daniel Korzekwa, Daniel Saez-Trigueros, David McHardy, Kamil Pokora, Jakub Lachowicz, Jaime Lorenzo-Trueba, and Viacheslav Klimkov. Enhancing audio quality for expressive neural text-to-speech, 2021.

[36] Cheng-I Jeff Lai, Erica Cooper, Yang Zhang, Shiyu Chang, Kaizhi Qian, Yi-Lun Liao, Yung-Sung Chuang, Alexander H. Liu, Junichi Yamagishi, David Cox, and James Glass. On the interplay between sparsity, naturalness, intelligibility, and prosody in speech synthesis, 2021.

[37] Raul Fernandez, David Haws, Guy Lorberbom, Slava Shechtman, and Alexander Sorin. Transplantation of conversational speaking style with interjections in sequence-to-sequence speech synthesis, 2022.

[38] Rowel Atienza. Efficientspeech: An on-device text to speech model, 2023.

[39] Yongqi Wang and Zhou Zhao. Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis, 2022.

[40] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Controllable accented text-to-speech synthesis, 2022.

[41] Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024.

[42] Chunhui Wang, Chang Zeng, Bowen Zhang, Ziyang Ma, Yefan Zhu, Zifeng Cai, Jian Zhao, Zhonglin Jiang, and Yong Chen. Ham-tts: Hierarchical acoustic modeling for token-based zero-shot text-to-speech with model and data scaling, 2024.

[43] Shivam Mehta, Ambika Kirkland, Harm Lameris, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Overflow: Putting flows on top of neural transducers for better tts, 2023.

[44] Mengxiao Bi, Heng Lu, Shiliang Zhang, Ming Lei, and Zhijie Yan. Deep feed-forward sequential memory networks for speech synthesis, 2018.

[45] Shivam Mehta, Éva Székely, Jonas Beskow, and Gustav Eje Henter. Neural hmms are all you need (for high-quality attention-free tts), 2022.

[46] Weidong Chen, Shan Yang, Guangzhi Li, and Xixin Wu. Drawspeech: Expressive speech synthesis using prosodic sketches as control conditions, 2025.

[47] Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie. Conversational end-to-end tts for voice agent, 2020.

[48] Paarth Neekhara, Jason Li, and Boris Ginsburg. Adapting tts models for new speakers using transfer learning, 2022.

[49] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Naturalspeech: End-to-end text to speech synthesis with human-level quality, 2022.

[50] Qing He, Zhiping Xiu, Thilo Koehler, and Jilong Wu. Multi-rate attention architecture for fast streamable text-to-speech spectrum modeling, 2021.

[51] Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka. Wavecyclegan: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks, 2018.

[52] Taejun Bak, Jae-Sung Bae, Hanbin Bae, Young-Ik Kim, and Hoon-Young Cho. Fastpitchformant: Source-filter based decomposed modeling for speech synthesis, 2021.

[53] CHEN Yushen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Kai Yu, Xie Chen, et al. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching.

[54] Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. Png bert: Augmented bert on phonemes and graphemes for neural tts, 2021.

[55] Tuomo Raitio, Javier Latorre, Andrea Davis, Tuuli Morrill, and Ladan Golipour. Improving the quality of neural tts using long-form content and multi-speaker multi-style modeling, 2023.

[56] Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. What the future brings: Investigating the impact of lookahead for incremental neural tts, 2020.

[57] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, 2020.

[58] Ron J. Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P. Kingma. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis, 2021.

[59] Zhengxi Liu, Qiao Tian, Chenxu Hu, Xudong Liu, Menglin Wu, Yuping Wang, Hang Zhao, and Yuxuan Wang. Controllable and lossless non-autoregressive end-to-end text-to-speech, 2022.

[60] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band melgan: Faster waveform generation for high-quality text-to-speech, 2020.

[61] Yeunju Choi, Youngmoon Jung, Youngjoo Suh, and Hoirin Kim. Learning to maximize speech quality directly using mos prediction for neural text-to-speech, 2022.

[62] Josef Schlittenlacher and Thomas Baer. Text-to-speech for the hearing impaired, 2021.

[63] Yanqing Liu, Ruiqing Xue, Lei He, Xu Tan, and Sheng Zhao. Delightfultts 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders, 2022.

[64] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi. A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis, 2018.

[65] Ankit Sharma, Puneet Kumar, Vikas Maddukuri, Nagasai Madamshettib, Kishore KG, Sahit Sai Sriram Kavurub, Balasubramanian Raman, and Partha Pratim Roy. Fast griffin lim based waveform generation strategy for text-to-speech synthesis, 2020.

[66] Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku. Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks, 2018.

[67] Yurii Rebryk and Stanislav Beliaev. Convoice: Real-time zero-shot voice style transfer with convolutional network, 2020.

[68] Yusuke Nakai, Yuki Saito, Kenta Udagawa, and Hiroshi Saruwatari. Multi-task adversarial training algorithm for multi-speaker neural text-to-speech, 2022.

[69] Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. Expressive text-to-speech using style tag, 2022.

[70] Yerin Choi and Myoung-Wan Koo. Dc comix tts: An end-to-end expressive tts with discrete code collaborated with mixer, 2023.

[71] Anandaswarup Vadapalli. An investigation of phrase break prediction in an end-to-end tts system, 2025.

[72] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba. Low-resource expressive text-to-speech using data augmentation, 2021.

[73] Elijah Gutierrez, Pilar Oplustil-Gallegos, and Catherine Lai. Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm, 2021.

[74] Chun-Yi Kuan, Chen An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo yiin Chang, and Hung yi Lee. Towards general-purpose text-instruction-guided voice conversion, 2024.

[75] Chung-Ming Chien and Hung yi Lee. Hierarchical prosody modeling for non-autoregressive speech synthesis, 2021.

[76] Wen-Chin Huang, Tomoki Hayashi, Xinjian Li, Shinji Watanabe, and Tomoki Toda. On prosody modeling for asr+tts based voice conversion, 2021.

[77] Zack Hodari, Oliver Watts, and Simon King. Using generative modelling to produce varied intonation for speech synthesis, 2019.

[78] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Expressive tts training with frame and style reconstruction loss, 2021.

[79] Liping Chen, Yan Deng, Xi Wang, Frank K. Soong, and Lei He. Speech bert embedding for improving prosody in neural tts, 2021.

[80] Slava Shechtman and Alex Sorin. Sequence to sequence neural speech synthesis with prosody modification capabilities, 2019.

[81] Zelin Ying, Chen Li, Yu Dong, Qiuqiang Kong, Qiao Tian, Yuanyuan Huo, and Yuxuan Wang. A unified front-end framework for english text-to-speech synthesis, 2024.

[82] Podakanti Satyajith Chary. Prosodic parameter manipulation in tts generated speech for controlled speech generation, 2024.

[83] Tuomo Raitio, Jiangchuan Li, and Shreyas Seshadri. Hierarchical prosody modeling and control in non-autoregressive parallel neural tts, 2022.

[84] Jing-Xuan Zhang, Li-Juan Liu, Yan-Nian Chen, Ya-Jun Hu, Yuan Jiang, Zhen-Hua Ling, and Li-Rong Dai. Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer, 2020.

[85] Sijing Chen, Yuan Feng, Laipeng He, Tianwei He, Wendi He, Yanni Hu, Bin Lin, Yiting Lin, Yu Pan, Pengfei Tan, Chengwei Tian, Chen Wang, Zhicheng Wang, Ruoye Xie, Jixun Yao, Quanlei Yan, Yuguang Yang, Jianhao Ye, Jingjing Yin, Yanzhen Yu, Huimin Zhang, Xiang Zhang, Guangcheng Zhao, Hongbin Zhou, and Pengpeng Zou. Takin: A cohort of superior quality zero-shot speech generation models, 2024.

[86] Phonetic and prosodic features f.

[87] Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Yufei Xia, Jinzhu Li, Sheng Zhao, Jinyu Li, and Naoyuki Kanda. An investigation of noise robustness for flow-matching-based zero-shot tts, 2024.

[88] Takaaki Saeki, Soumi Maiti, Xinjian Li, Shinji Watanabe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining, 2023.

[89] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, 2023.

[90] Anders R. Bargum, Stefania Serafin, and Cumhur Erkut. Reimagining speech: A scoping review of deep learning-powered voice conversion, 2023.

[91] Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment, 2024.

[92] Taejun Bak, Youngsik Eom, SeungJae Choi, and Young-Sun Joo. Multiverse: Efficient and expressive zero-shot multi-task text-to-speech, 2024.

[93] Hieu-Thi Luong and Junichi Yamagishi. Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech, 2019.

[94] Kenichi Fujita, Takanori Ashihara, Marc Delcroix, and Yusuke Ijima. Lightweight zero-shot text-to-speech with mixture of adapters, 2024.

[95] Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. Make-a-voice: Unified voice synthesis with discrete representation, 2023.

[96] Yinlin Guo, Yening Lv, Jinqiao Dou, Yan Zhang, and Yuehai Wang. Fly-tts: Fast, lightweight and high-quality end-to-end text-to-speech synthesis. *arXiv preprint arXiv:2407.00753*, 2024.

[97] Yifan Yang, Ziyang Ma, Shujie Liu, Jinyu Li, Hui Wang, Lingwei Meng, Haiyang Sun, Yuzhe Liang, Ruiyang Xu, Yuxuan Hu, et al. Interleaved speech-text language models are simple streaming text to speech synthesizers. *arXiv preprint arXiv:2412.16102*, 2024.

[98] Nikolaos Ellinas, Georgios Vamvoukakis, Konstantinos Markopoulos, Aimilios Chalamandaris, Georgia Maniati, Panos Kakoulidis, Spyros Raptis, June Sig Sung, Hyoungmin Park, and Pirros Tsiakoulis. High quality streaming speech synthesis with low, sentence-length-independent latency, 2021.

[99] Trung Dang, David Aponte, Dung Tran, Tianyi Chen, and Kazuhito Koishida. Zero-shot text-to-speech from continuous text streams, 2024.

[100] Rohan Badlani, Adrian Łancucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. One tts alignment to rule them all, 2021.

[101] Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe. Espnet2-tts: Extending the edge of tts research, 2021.

[102] Mingyang Zhang, Yi Zhou, Li Zhao, and Haizhou Li. Transfer learning from speech synthesis to voice conversion with non-parallel training data, 2021.

[103] Hy Quy Nguyen, Siu Wa Lee, Xiaohai Tian, Minghui Dong, and Eng Siong Chng. High quality voice conversion using prosodic and high-resolution spectral features, 2015.

[104] Yinghao Aaron Li, Cong Han, and Nima Mesgarani. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models, 2022.

[105] Anders R Bargum, Stefania Serafin, and Cumhur Erkut. Reimagining speech: A scoping review of deep learning-powered voice conversion. *arXiv preprint arXiv:2311.08104*, 2023.

[106] Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Dacheng Yin, Yucheng Zhao, and Wenjun Zeng. Zero-shot text-to-speech for text-based insertion in audio narration, 2021.

[107] Haogeng Liu, Tao Wang, Ruibo Fu, Jiangyan Yi, Zhengqi Wen, and Jianhua Tao. Unifyspeech: A unified framework for zero-shot text-to-speech and voice conversion, 2023.

[108] Jiatong Shi, Yun Tang, Ann Lee, Hirofumi Inaguma, Changhan Wang, Juan Pino, and Shinji Watanabe. Enhancing speech-to-speech translation with multiple tts targets, 2023.

[109] Javier Latorre, Charlotte Bailleul, Tuuli Morrill, Alistair Conkie, and Yannis Stylianou. Combining speakers of multiple languages to improve quality of neural voices, 2021.

[110] Jee weon Jung, Wangyou Zhang, Soumi Maiti, Yihan Wu, Xin Wang, Ji-Hoon Kim, Yuta Matsunaga, Seyun Um, Jinchuan Tian, Hye jin Shim, Nicholas Evans, Joon Son Chung, Shinnosuke Takamichi, and Shinji Watanabe. Text-to-speech synthesis in the wild, 2024.

[111] Marcel de Korte, Jaebok Kim, and Esther Klabbers. Efficient neural speech synthesis for low-resource languages through multilingual modeling, 2020.

[112] Kohki Mametani, Tsuneo Kato, and Seiichi Yamamoto. Investigating context features hidden in end-to-end tts, 2019.

[113] Julia Kaiwen Lau, Kelvin Kai Wen Kong, Julian Hao Yong, Per Hoong Tan, Zhou Yang, Zi Qian Yong, Joshua Chern Wey Low, Chun Yong Chong, Mei Kuan Lim, and David Lo. Synthesizing speech test cases with text-to-speech? an empirical study on the false alarms in automated speech recognition testing, 2023.

[114] Alexandra Vioni, Georgia Maniati, Nikolaos Ellinas, June Sig Sung, Inchul Hwang, Aimilios Chalamandaris, and Pirros Tsiakoulis. Investigating content-aware neural text-to-speech mos prediction using prosodic and linguistic features, 2023.

[115] Jan Chorowski, Ron J. Weiss, Rif A. Saurous, and Samy Bengio. On using backpropagation for speech texture generation and voice conversion, 2018.

[116] Christoph Minixhofer, Ondřej Klejch, and Peter Bell. Ttsds – text-to-speech distribution score, 2024.

[117] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet, 2019.

[118] Eunwoo Song, Kyungguen Byun, and Hong-Goo Kang. Excitnet vocoder: A neural excitation model for parametric speech synthesis systems, 2019.

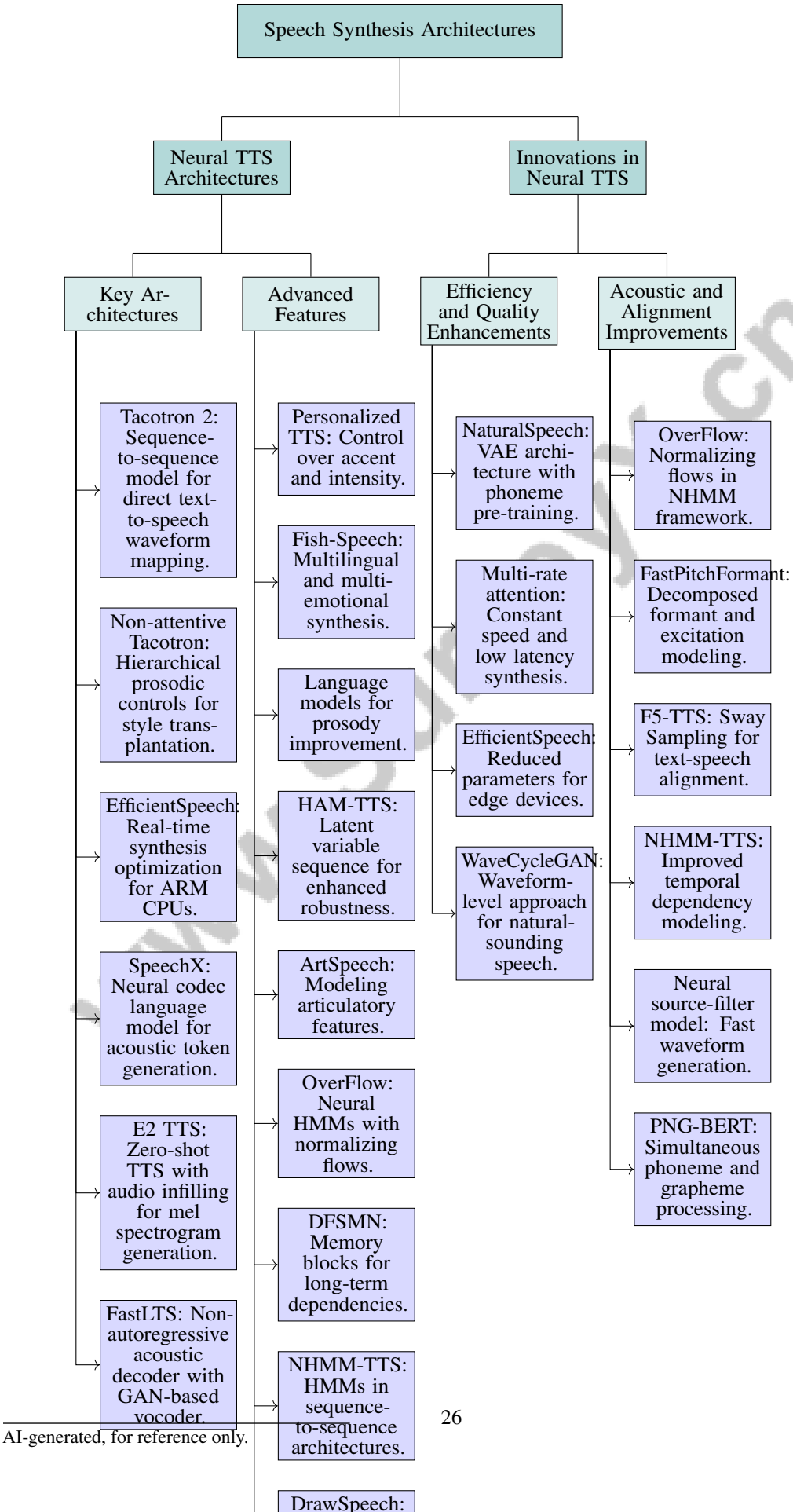
[119] Seongyeon Park, Bohyung Kim, and Tae hyun Oh. Automatic tuning of loss trade-offs without hyper-parameter search in end-to-end zero-shot speech synthesis, 2023.

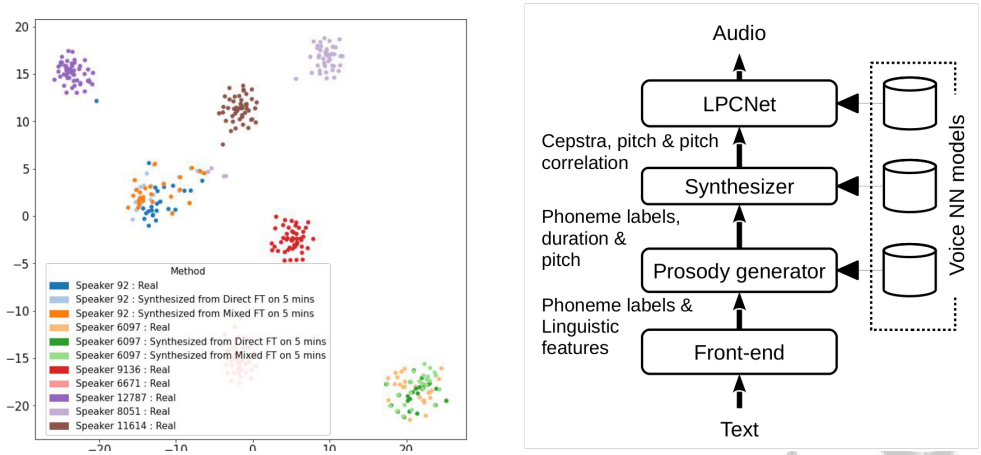
[120] Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech, 2023.

[121] Quanxiu Wang, Hui Huang, Mingjie Wang, Yong Dai, Jinzuomu Zhong, and Benlai Tang. Prior-agnostic multi-scale contrastive text-audio pre-training for parallelized tts frontend modeling, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.





(a) Speaker Classification Using Deep Learning and Audio Features[48]

(b) A diagram illustrating a speech synthesis process[26]

Figure 3: Examples of Neural TTS Architectures

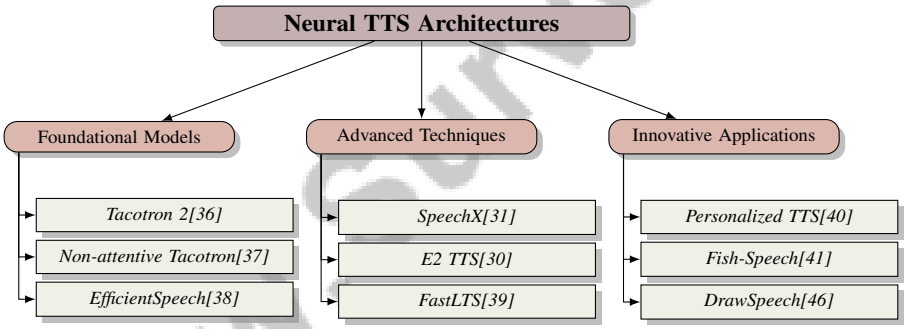
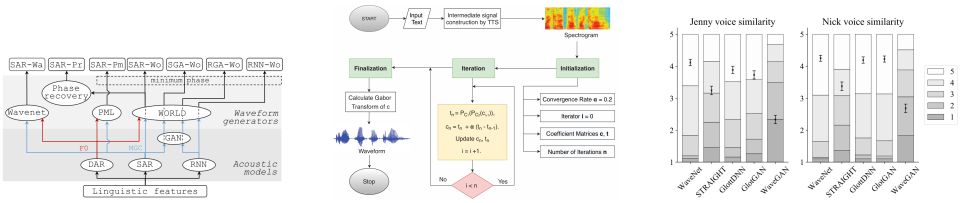


Figure 4: This figure illustrates the hierarchical categorization of Neural TTS Architectures into foundational models, advanced techniques, and innovative applications, showcasing significant advancements and applications in the field.



(a) Waveform synthesis and phase recovery system[64]

(b) A Flowchart of a Text-to-Speech (TTS) System[65]

(c) Voice Similarity Comparison Across Different Generative Models[66]

Figure 5: Examples of Techniques for Waveform Generation

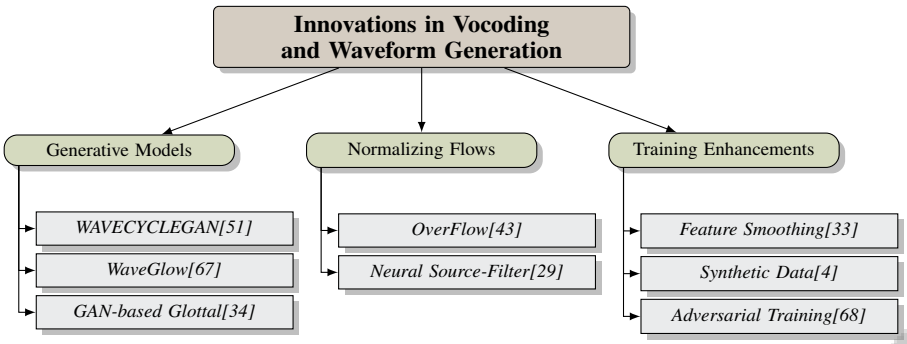


Figure 6: This figure illustrates recent innovations in vocoding and waveform generation, categorizing them into generative models, normalizing flows, and training enhancements, showcasing advancements in synthesized speech quality and efficiency.

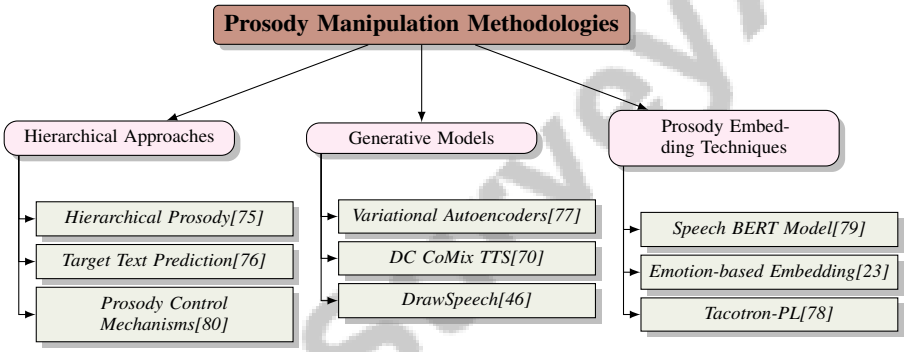


Figure 7: This figure illustrates key methodologies for prosody manipulation in TTS systems, categorized into hierarchical approaches, generative models, and prosody embedding techniques. Each category showcases different methods aimed at enhancing speech expressiveness and naturalness.

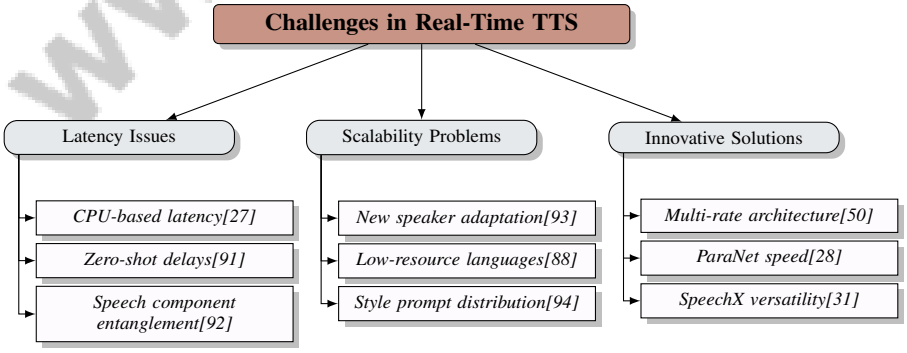


Figure 8: This figure illustrates the challenges faced by real-time Text-to-Speech (TTS) systems, categorized into latency issues, scalability problems, and innovative solutions. Each category highlights specific aspects such as CPU-based latency, adaptation to new speakers, and advanced architectures like multi-rate attention and ParaNet for improved performance.