Advanced Concepts in Text-to-Speech Synthesis: A Survey

www.surveyx.cn

Abstract

This survey paper explores the advanced concepts and technologies in text-to-speech (TTS) synthesis, emphasizing innovations in artificial intelligence and machine learning. Key technologies such as discrete representation, vector quantization, and codebook learning are examined, highlighting their role in enhancing speech synthesis quality and efficiency. The paper delves into models like VALL-E and VQ-VAE, which utilize multi-level discrete modeling and neural codecs to enable high-quality, end-to-end speech synthesis. Advancements in neural networks, including sequence-to-sequence frameworks and generative adversarial networks, are discussed for their contributions to improving synthesis speed and naturalness. The integration of large language models with neural codecs is also explored, showcasing their synergistic effects in refining TTS systems. Applications of these technologies span multilingual communication, accessibility, and personalized speech synthesis. The paper concludes by addressing future research directions, emphasizing the potential for further innovations to enhance the adaptability and scalability of TTS systems across diverse applications. These advancements promise to drive significant progress in AI and machine learning, transforming the landscape of speech synthesis.

1 Introduction

1.1 Significance of TTS Synthesis

Text-to-speech (TTS) synthesis is a fundamental technology in artificial intelligence, vital for enhancing communication and accessibility across various applications. Its ability to generate high-quality, natural-sounding speech is essential for virtual assistants and accessibility tools that demand clear audio outputs [1]. TTS systems are particularly impactful in low-resource and multilingual settings, effectively bridging linguistic divides and enabling the production of human-like speech [2].

Recent advancements in TTS have overcome the limitations of traditional phoneme-level duration models, which often resulted in inferior audio quality. Innovations have led to robust zero-shot synthesis, alleviating the constraints of previous methods reliant on speaker adaptation and intricate feature engineering [3]. Moreover, TTS synthesis enhances the spoken input and output capabilities of large language models (LLMs), facilitating more natural human-machine interactions [4].

Beyond applications in virtual assistants and accessibility, TTS synthesis is crucial for voice conversion (VC) systems, which transform speech from a source speaker to mimic a target speaker while maintaining linguistic integrity [5]. This capability is vital for personalizing speech characteristics, thereby enhancing user engagement. Additionally, TTS systems are increasingly used to generate sound effects from text prompts, showcasing their versatility in gaming and virtual reality.

The challenges faced by TTS systems, such as processing complex linguistic features and managing polyphonic expressions, underscore the need for ongoing research and development. Despite these challenges, TTS synthesis remains a cornerstone technology for advancing AI applications, evolving

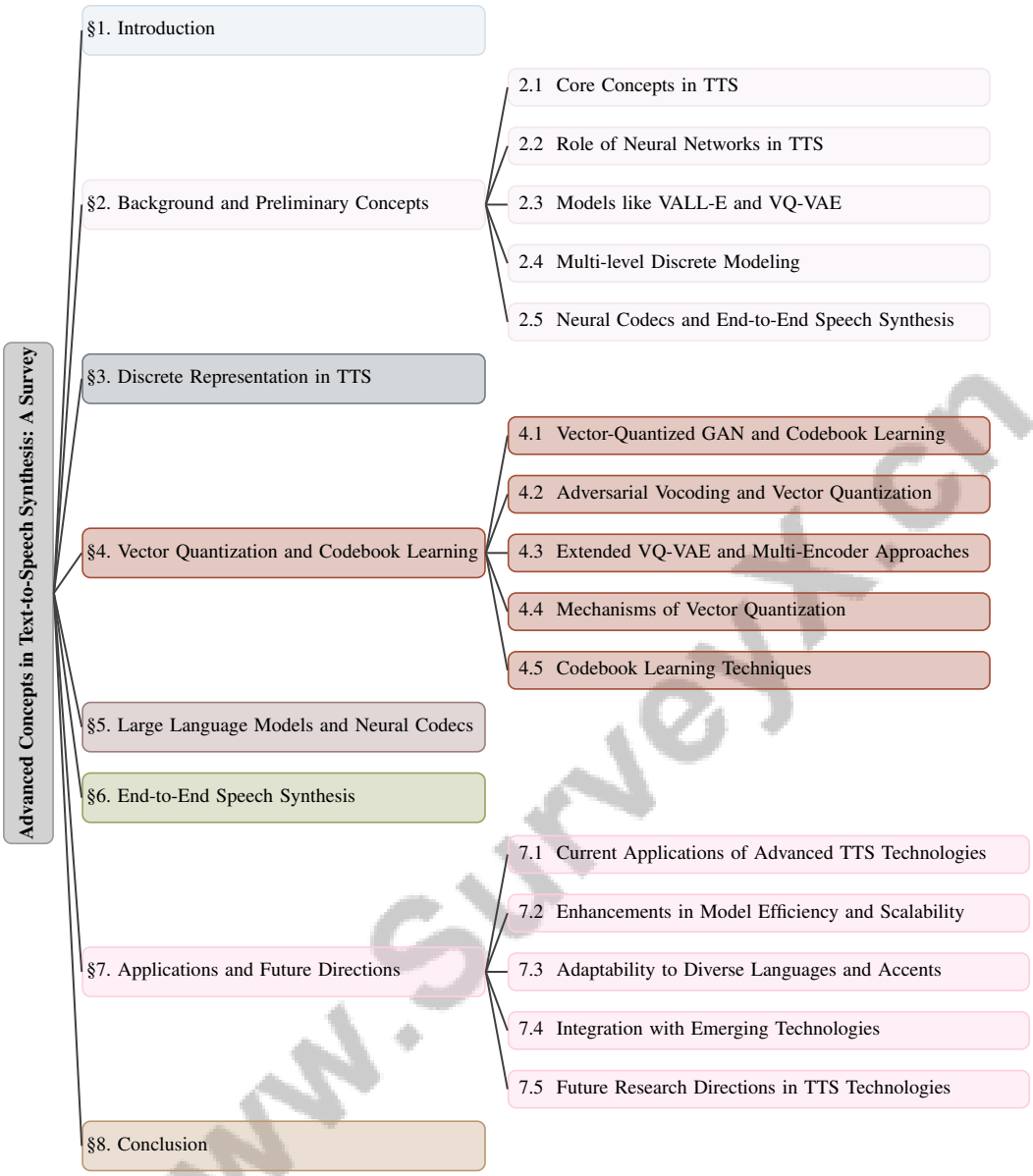


Figure 1: chapter structure

to meet the demands of modern communication technologies. The significance of TTS synthesis is further emphasized by its role in improving the quality and naturalness of synthetic speech through advanced algorithms and techniques, including neural network models and TTS-driven data augmentation methods, which collectively enhance intelligibility and emotional expressiveness. As a result, TTS synthesis is integral in various sectors, including accessibility tools, virtual assistants, and navigation systems, facilitating more engaging and lifelike user experiences [6, 7, 8].

1.2 Advancements in Neural Networks

Advancements in neural networks have significantly enhanced TTS systems, primarily through the adoption of innovative architectures and learning paradigms. Sequence-to-sequence (Seq2Seq) learning frameworks have shown promise in voice conversion tasks, eliminating the need for parallel data and enabling more flexible TTS processes [9]. The neural source-filter model has demonstrated superior generation speed compared to traditional autoregressive models while maintaining comparable speech quality, thus improving TTS efficiency [3].

Incorporating vector quantized variational autoencoders (VQ-VAE) within frameworks like DiscreTalk has streamlined the end-to-end TTS process by removing human-designed acoustic features, thereby enhancing speech synthesis quality [10]. Models such as SoCodec address complexity and efficiency challenges in language model-based TTS approaches by introducing shorter, multi-stream discrete semantic sequences, optimizing performance [11].

Exploration of autoregressive diffusion transformers (ARDiT) has further advanced TTS systems by enabling the generation of continuous audio tokens, resulting in higher fidelity audio generation and improved performance in text-to-speech tasks [12]. The use of generative adversarial networks (GANs) has been pivotal in addressing the limitations of deep neural networks in capturing the stochastic variation of glottal waveforms, advancing the naturalness and variability of synthesized speech [13].

Additionally, integrating high-capacity VC models with low-capacity, single-speaker, monolingual acoustic models has enhanced TTS performance, particularly in cross-lingual scenarios [4]. Recognizing prosodic elements within speech signals has informed the development of TTS systems capable of producing more natural and intelligible audio, as evidenced by ongoing research [14].

These advancements underscore the transformative role of neural networks in TTS technology, driving improvements in speech quality, diversity, and expressiveness. As research progresses, innovations will likely focus on enhancing the naturalness and emotional expressiveness of synthesized speech, integrating high-quality, language-specific datasets, and optimizing model pre-training techniques to ensure better performance for target speakers. Such improvements will facilitate more effective communication in diverse applications, including accessibility tools and virtual assistants, while addressing challenges in multilingual support and the need for high-quality training data essential for robust TTS systems [6, 15, 16].

1.3 Key Technologies and Models

In TTS synthesis, discrete representation and vector quantization have emerged as transformative technologies, significantly enhancing the quality and efficiency of speech generation systems. Vector-quantized autoencoders exemplify this advancement, utilizing adversarial training to refine intermediate speech representations, thereby improving the fidelity of synthesized speech [17]. This highlights the critical role of vector quantization in effectively encoding acoustic information essential for generating high-quality audio outputs.

Wave-Tacotron exemplifies innovation in TTS by eliminating fixed intermediate representations, streamlining the synthesis pipeline and improving efficiency [18]. Similarly, ParaNet's non-autoregressive architecture facilitates rapid conversion of text to mel spectrograms, significantly boosting synthesis speed and demonstrating the potential of novel architectures in TTS systems [19].

The integration of pretrained self-supervised speech representations, as seen in WavThruVec, underscores advancements in acoustic representations to enhance TTS quality [20]. The Language-Codec model employs a Masked Channel Residual Vector Quantization mechanism, optimizing the codec structure for improved compatibility with speech language models, thus advancing TTS adaptability [21].

Innovations like multi-band MelGAN improve upon traditional methods by incorporating multi-band processing and utilizing a multi-resolution STFT loss, which enhances waveform generation speed and quality [22]. ZMM-TTS introduces discrete speech representations that disentangle speaker and language information, showcasing the versatility of discrete modeling techniques in multilingual and multispeaker scenarios [23].

The development of VAENAR-TTS, synthesizing speech without requiring phoneme-level durations, exemplifies advancements in naturalness through variational autoencoder-based techniques [24]. Collectively, these technologies and models illustrate the transformative impact of discrete representation and vector quantization in TTS systems, driving significant improvements in speech synthesis quality, efficiency, and diversity across applications. As these technologies evolve, they promise to enhance the scalability and adaptability of TTS systems, enabling integration into diverse fields such as accessibility, entertainment, and education [6]. The taxonomy of neural TTS methods introduced by Tan et al. provides a structured framework for understanding these advancements,

focusing on key components such as text analysis, acoustic models, and vocoders, with an emphasis on end-to-end models [8].

1.4 Overview of Paper Structure

This survey is meticulously designed to provide an in-depth examination of advanced concepts and technologies in TTS synthesis, with a focus on neural network architectures and end-to-end speech generation models. It addresses key components such as text analysis, acoustic modeling, and vocoding, while exploring cutting-edge topics like fast TTS, low-resource TTS, robust TTS, expressive TTS, and adaptive TTS. The survey highlights the implications of TTS technology across various sectors, including accessibility and virtual assistance, and discusses future research directions and available resources, making it a valuable resource for both academic researchers and industry practitioners [6, 8].

The paper begins with an **Introduction**, emphasizing the significance of TTS synthesis in AI and machine learning, followed by advancements in neural networks that have enabled high-quality audio generation. Key technologies and models such as discrete representation and vector quantization are introduced, setting the stage for subsequent analyses.

In **Section 2**, the paper delves into the **Background and Preliminary Concepts**, defining core concepts like discrete representation, vector quantization, and codebook learning. This section discusses the role of neural networks in TTS, focusing on models like VALL-E and VQ-VAE, and introduces multi-level discrete modeling and neural codecs, explaining their relevance to end-to-end speech synthesis.

Section 3 explores **Discrete Representation in TTS**, examining how vector quantization and codebook learning contribute to efficient and high-quality speech synthesis. The section provides examples of models utilizing these techniques, such as DelightfulTTS 2, which integrates a codec network with an acoustic model for enhanced synthesis [17].

Section 4 focuses on **Vector Quantization and Codebook Learning**, detailing the mechanisms and applications of these techniques in neural networks. The benefits and challenges associated with these methods are discussed, referencing models like Enhanced Neural Text-to-Speech (ENTTS), which employs GAN-based acoustic modeling and VAE conditioning for high-quality audio generation [25].

In **Section 5**, the integration of **Large Language Models (LLMs)** and **Neural Codecs** is explored in detail, specifically focusing on the model VALL-E, which utilizes discrete codes from a neural audio codec to treat TTS synthesis as a conditional language modeling task. This innovative approach, trained on an extensive dataset of 60,000 hours of English speech, enables VALL-E to generate high-quality, personalized speech with remarkable speaker similarity and the ability to preserve emotional tone and environmental acoustics from brief audio prompts. The section discusses empirical studies assessing various integration techniques between LLMs and VALL-E, revealing that coupling LLMs as text encoders with VALL-E yields superior performance in speech synthesis, achieving significant improvements in both speaker similarity and word error rate compared to traditional models [26, 27].

Section 6 discusses **End-to-End Speech Synthesis**, exploring the concept and advantages of this approach over traditional methods. Insights into the performance and applications of these systems are provided, drawing on the structured methodology of models like Tacotron, which offers a detailed framework for end-to-end TTS [28].

In **Section 7**, titled **Applications and Future Directions**, the paper analyzes the current applications of advanced TTS technologies across various sectors, such as accessibility tools, navigation systems, and virtual assistants. It discusses potential developments and research avenues in TTS, including enhancements in naturalness, multilingual support, emotional expression, and the application of deep learning techniques. The section highlights the significance of neural network-based approaches, challenges in TTS synthesis, and implications for future innovations in the field [15, 29, 8, 6, 30]. The adaptability of TTS models to diverse languages and accents is analyzed, along with the integration of TTS technologies with emerging technologies.

Finally, the **Conclusion** summarizes key points discussed throughout the paper, reinforcing the importance of discrete representation, vector quantization, and neural codecs in advancing TTS technology. The potential impact of these advancements on future AI and machine learning applications

is highlighted, drawing on the structured approach exemplified by the Voice Transformer Network (VTN) [31].The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Core Concepts in TTS

Text-to-speech (TTS) synthesis hinges on discrete representation, vector quantization (VQ), and codebook learning to enhance speech generation systems' quality and efficiency. Discrete representation captures essential speech characteristics, facilitating high-quality, human-like audio synthesis [8, 23]. This approach is exemplified by discrete diffusion models capable of generating sound from textual descriptions [32].

Vector quantization encodes acoustic signals into discrete tokens, transforming them into audio waveforms while minimizing redundancy [33, 34]. Techniques like group residual vector quantization ensure high reconstruction quality with fewer codebooks [35]. By encoding complex speech features into discrete embeddings, VQ effectively supports audio synthesis from text [36].

Codebook learning refines intermediate speech representations, crucial in low-resource settings where paired text and speech data are limited [23, 37]. It separates content and speaker information, enhancing system adaptability across diverse linguistic contexts [38].

The challenge of compressing speech signals into concise token sequences stems from the rich information in speech, including phonetics, prosody, and speaker identity [11]. Discrete representation techniques are vital for generating natural, high-quality speech [9]. As TTS technologies evolve, the interplay among discrete representation, vector quantization, and codebook learning will enhance TTS system performance and versatility, facilitating applications in accessibility, voice conversion, and accent modification.

2.2 Role of Neural Networks in TTS

Neural networks have revolutionized text-to-speech (TTS) technologies, producing more natural-sounding speech than traditional rule-based systems, which often struggled with audio quality [39]. By integrating neural networks, TTS systems optimize the acoustic model and vocoder, reducing cascaded errors and improving synthesis quality [17].

A significant challenge in TTS is the monotonic nature of phoneme-to-audio alignment, where decoder-only transformer structures may lack accuracy [40]. Neural networks address these inefficiencies, enabling streamlined designs. Generative adversarial networks (GANs) enhance synthetic speech quality by minimizing divergence between natural and generated speech parameters [13].

Autoregressive TTS models often exhibit slow synthesis speeds due to their sequential operation, inefficient for modern parallel hardware [19]. Non-autoregressive architectures, such as Multi-band MelGAN, utilize neural networks to generate high-quality audio efficiently, enhancing real-time application viability [22].

Novel methods like Fish-Speech leverage a fast-slow Dual Autoregressive architecture to improve synthesis quality [38]. VoxInstruct exemplifies the role of neural networks in TTS by integrating content and descriptive information for expressive speech generation, showcasing end-to-end synthesis potential [41].

The reliance on large datasets for training poses challenges in low-resource scenarios. Neural networks provide solutions through transfer and self-supervised learning approaches, leveraging pre-existing models and data to enhance performance in resource-constrained environments [42]. For instance, ZMM-TTS employs self-supervised learning to utilize large amounts of unlabeled data, enhancing multilingual and multispeaker synthesis [23].

Neural networks also address inefficiencies in current waveform synthesis methods, which struggle to model distinct characteristics of voiced and unvoiced speech segments, often resulting in unnatural artifacts [43]. The integration of neural codec language modeling with multi-task learning, as seen in SpeechX, illustrates the potential of neural networks to improve TTS capabilities by enhancing synthesis quality and efficiency [44].

Neural networks continue to drive advancements in TTS technologies, improving synthesis quality, efficiency, and adaptability across diverse linguistic environments. As research progresses, the integration of neural network-based approaches is expected to refine and expand TTS systems, addressing existing challenges and paving the way for future innovations. Models like Text-Instruction-Guided Voice Conversion (TIG-VC) showcase the role of neural networks in advancing TTS technologies through efficient and high-quality voice conversion [5]. Additionally, large pre-trained models facilitate high-quality synthetic text generation, essential for producing natural-sounding speech [39].

2.3 Models like VALL-E and VQ-VAE

Models such as VALL-E and VQ-VAE have introduced innovative methodologies in text-to-speech (TTS) synthesis, significantly enhancing generated speech's quality and adaptability. VALL-E functions as a conditional language model that synthesizes speech by generating acoustic codes from phoneme sequences and acoustic prompts, enabling diverse and contextually appropriate outputs with zero-shot capabilities. Its architecture supports speaker-language disentanglement, a critical innovation that separates these components, traditionally managed by a single model in standard polyglot approaches [4]. VALL-E R refines this model through a phoneme monotonic alignment strategy and codec-merging technique, enhancing TTS synthesis robustness and efficiency. VALL-E 2 further advances performance with repetition-aware sampling and grouped code modeling.

Conversely, VQ-VAE employs vector quantization to discretize latent acoustic features, effectively separating phonetic content from speaker identity. This decoupling enhances control over speech attributes, improving prosody and naturalness in synthesized audio. The model's use of multiple codebooks and multi-stage representation marks a significant advancement over traditional single-codebook methods, exemplified in applications like VQVC+, which integrates a U-Net architecture with vector quantization to enhance audio quality in voice conversion tasks, transforming a source speaker's timbre, accent, and tone while preserving original linguistic content, even in one-shot scenarios. The vector quantization technique improves disentanglement of speaker and content features, essential for achieving high audio naturalness and speaker similarity in generated speech [45, 46, 47, 35]. Such innovations are particularly beneficial in multilingual speech synthesis, addressing challenges associated with diverse linguistic contexts.

The integration of advanced neural architectures complements the capabilities of models like VQ-VAE. For instance, the ARDiT method generates audio without requiring discrete tokenization, allowing for high-bitrate continuous speech representation that enhances reconstruction accuracy and generative capabilities [12]. This underscores the potential for further advancements in neural architectures, improving TTS system stability and efficiency.

As research progresses, models like VALL-E and VQ-VAE are expected to undergo refinements, addressing existing challenges and broadening their applications across various domains. Recent advancements in TTS underscore the critical role of discrete latent spaces and advanced neural architectures in improving flexibility, efficiency, and overall performance. By leveraging discrete speech tokens and sophisticated models, researchers have demonstrated enhanced capabilities in generating diverse and natural speech outputs while tackling challenges such as intelligibility and speaker consistency. These innovations pave the way for robust TTS applications and open new avenues for integrating multimodal tasks and improving data transfer efficiency, reshaping the landscape of speech synthesis for a new era of expressive and context-aware TTS technologies [28, 8, 48, 49, 50].

2.4 Multi-level Discrete Modeling

Multi-level discrete modeling represents a transformative approach in text-to-speech (TTS) synthesis, employing a structured framework to capture intricate speech characteristics across various linguistic and prosodic dimensions. This method utilizes advanced discrete speech tokenization and large language models, enabling the generation of diverse, natural-sounding speech while addressing challenges like intelligibility and speaker consistency. By integrating semantic and acoustic tokens within a two-stage framework, it improves the alignment of text with speech, enhancing overall quality and expressiveness, particularly beneficial in applications such as accessibility tools and virtual assistants [51, 8, 48, 6, 49]. This modeling technique facilitates high-quality speech synthesis

by leveraging multiple discrete levels to encode and decode speech features, enhancing TTS systems' expressiveness and adaptability.

A key innovation in this domain is the incorporation of monotonic alignment constraints into decoder-only Transformer models, as demonstrated by VALL-T. This approach enhances TTS systems' robustness and controllability, ensuring generated speech aligns accurately with input text while maintaining natural prosodic variations [52]. By addressing inherent alignment challenges in TTS tasks, VALL-T improves synthesis quality and reliability.

The CHiVE model introduces a dynamic hierarchical structure adapting to linguistic inputs, creating a meaningful prosodic space for feature sampling. This hierarchical approach allows varied prosodic patterns, contributing to more expressive and natural-sounding speech synthesis [53]. The ability to adjust prosodic features dynamically based on linguistic context is crucial for achieving high-quality TTS outputs.

Additionally, combining a speaker-aware text encoder with a VALL-E based acoustic decoder exemplifies multi-level discrete modeling's potential in adapting to unseen speakers. This innovation facilitates better speaker adaptation, particularly relevant in zero-shot scenarios where the system generates speech for speakers not present in the training data [54]. Such adaptability is essential for expanding TTS applications across diverse speaker profiles and languages.

The Gaussian Inverse Autoregressive Flow (IAF) simplifies training by providing a closed-form computation for KL divergence, enabling efficient distillation and faster inference [55]. This advancement contributes to the scalability and efficiency of multi-level discrete models, making them more practical for real-world applications.

Future research in multi-level discrete modeling should focus on developing data-efficient TTS models, enhancing expressiveness, and addressing challenges in synthesizing speech for low-resource languages [8]. As these models continue to evolve, they promise to refine TTS systems' capabilities, enabling nuanced and contextually appropriate speech synthesis across various applications.

2.5 Neural Codecs and End-to-End Speech Synthesis

Neural codecs are pivotal in advancing end-to-end text-to-speech (TTS) synthesis by facilitating a seamless transition from text to high-quality audio, eliminating the need for traditional intermediate representations and enhancing synthesis efficiency and fidelity. The integration of neural codecs in TTS systems is exemplified by models like VQMIVC, which employs content and speaker encoders, a pitch extractor, and a decoder for voice conversion, showcasing versatility in synthesis and conversion tasks [56].

Adversarial vocoding techniques further enhance neural codec capabilities. Frameworks like EATS utilize generative adversarial networks (GANs) to directly map sequences to audio waveforms, improving the naturalness and variability of synthesized speech. This method addresses traditional vocoding limitations, enhancing TTS output quality. Model compression strategies, such as sparsity and quantization, are crucial for optimizing neural codecs like the WaveNet vocoder, enabling faster and more resource-efficient synthesis [57].

Recent innovations in neural audio coding, such as the Low Frame-rate Speech Codec (LFSC), highlight the potential of neural codecs to improve TTS systems' quality and efficiency. LFSC is designed for efficient training and inference of Speech LLMs, compressing audio at a frame rate of 21.5 FPS and a bitrate of 1.89 kbps [58]. Similarly, the Efficient Speech Codec (ESC) employs a transformer-based architecture with cross-scale residual vector quantization for effective speech signal compression [59]. The X-Codec combines semantic features from a pre-trained semantic encoder with acoustic features and utilizes residual vector quantization (RVQ) for tokenization, incorporating a semantic reconstruction loss to enhance synthesized speech quality [60].

Moreover, the NDVQ framework improves audio synthesis robustness by incorporating learnable variance into the quantization process, enhancing resilience to input data variations [47]. The Single-Codec model exemplifies neural codecs' efficiency by utilizing a single codebook for speech compression and reconstruction, optimizing the balance between data compression and audio quality [61]. The SALMONN-omni model demonstrates neural codecs' potential by providing full-duplex speech understanding and generation capabilities, processing input and output speech in real-time through a novel framework [62].

The TF-Codec, a low-latency neural speech codec, employs latent-domain predictive coding within the VQ-VAE framework to efficiently remove temporal redundancies, showcasing neural codecs' potential to optimize synthesis processes [63]. Additionally, the CWT-Vocoder utilizes Continuous Wavelet Transform for analyzing and decomposing speech features, enabling high-quality waveform generation [37]. The effectiveness of MQTTS highlights discrete representations' resilience to noise, capturing the diversity found in spontaneous speech [33].

These advancements not only enhance end-to-end TTS systems' performance but also pave the way for future innovations in speech synthesis, offering new research and application avenues in fields from accessibility to entertainment. As neural codecs evolve, their integration into TTS systems promises to elevate synthesized speech quality and efficiency, contributing to the ongoing advancement of artificial intelligence and machine learning technologies [39].

In recent years, advancements in text-to-speech (TTS) technology have significantly transformed the landscape of synthetic voice generation. A critical aspect of these advancements lies in the exploration of discrete representation methods, which have emerged as a focal point for enhancing synthesis quality and expressiveness. As detailed in Figure 2, the hierarchical structure of these discrete representation methods is illustrated, showcasing key components such as the dynamic quantized representation module, phoneme-level discrete latent representation, and WaveNet autoencoders. This figure not only delineates the frameworks and applications associated with these methods but also emphasizes the advancements that contribute to improved adaptability and overall synthesis quality. By integrating these elements, researchers can better understand the intricate interplay between various components in TTS systems, paving the way for future innovations in the field.

3 Discrete Representation in TTS

3.1 Dynamic Quantized Representation Module

The dynamic quantized representation module is pivotal in TTS systems, converting discrete representations into high-quality audio outputs. Frameworks like Token Transducer++ exemplify its functionality by transforming text into semantic tokens, which are then used to generate raw waveforms from acoustic tokens, ensuring high fidelity in expressive speech synthesis. VQ-VAE further showcases this module by encoding input data into latent representations and obtaining discrete latents through an embedding space, enhancing control over synthesized speech features [64].

The MSMC-VQ-VAE method illustrates the role of this module by encoding Mel spectrograms into multi-stage multi-codebook representations, optimizing synthesis, and minimizing reconstruction errors [34]. This capability is crucial for improving audio quality. HiFi-Codec uses group residual vector quantization, ensuring high-quality outputs with fewer codebooks [35].

Incorporating environmental and speaker factors into TTS systems enriches the synthesis process, as demonstrated by embedding extractors [65]. The ATTS2S-VC method enhances voice conversion tasks through attention and context preservation mechanisms, showcasing the module's capabilities [9].

Moreover, VoxInstruct generates both coarse-grained and detailed acoustic tokens, highlighting the application of discrete representation in TTS systems [41]. The ARDiT model leverages diffusion probabilistic models for high-quality audio generation in a single evaluation step, emphasizing the potential of dynamic quantized representation modules [12].

Figure 3 illustrates the hierarchical structure of the Dynamic Quantized Representation Module in TTS systems, categorizing key methods, applications, and capabilities that enhance audio quality and expressiveness. These modules are instrumental in advancing TTS systems by optimizing synthesis processes and enhancing audio quality and expressiveness. Their evolving capabilities promise to improve the adaptability and performance of TTS technologies, contributing to broader AI and machine learning applications. Training on diverse linguistic databases illustrates the module's adaptability in multilingual contexts [66].

3.2 Phoneme-level Discrete Latent Representation

Phoneme-level discrete latent representations are vital for enhancing TTS systems by disentangling prosodic information from linguistic content and speaker characteristics [67]. This approach enables

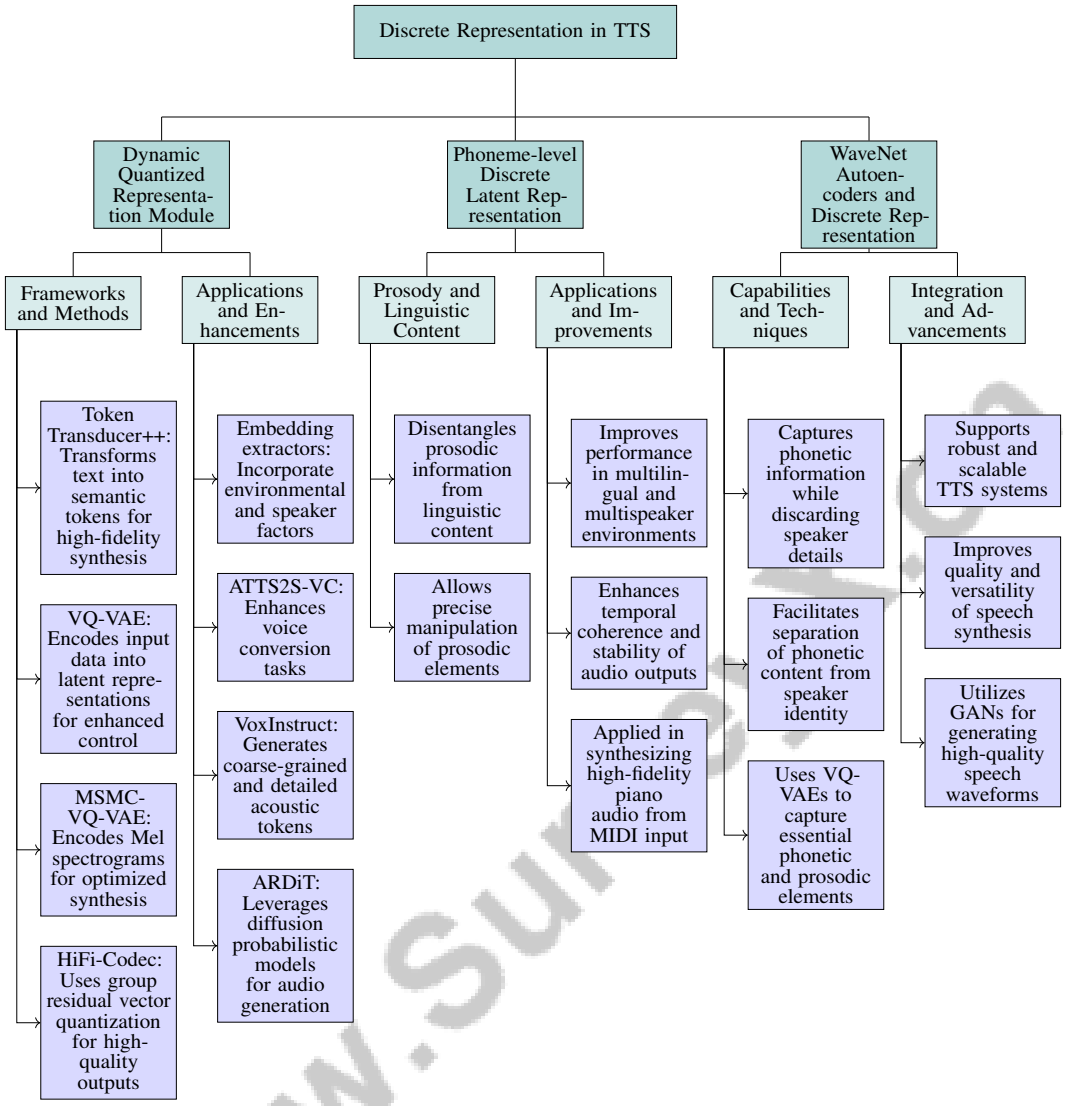


Figure 2: This figure illustrates the hierarchical structure of discrete representation methods in TTS, detailing the dynamic quantized representation module, phoneme-level discrete latent representation, and WaveNet autoencoders. It highlights the frameworks, applications, and advancements contributing to improved synthesis quality, adaptability, and expressiveness.

more natural and expressive speech by accurately capturing prosody nuances while preserving linguistic integrity. By focusing on phoneme-level features, these representations allow precise manipulation of prosodic elements such as intonation, stress, and rhythm.

This advancement, supported by discrete speech tokenization and large language models (LLMs), improves performance across tasks like speech recognition and speech-to-speech translation, allowing TTS systems to generate natural and contextually appropriate outputs [49, 48]. This capability is particularly beneficial in multilingual and multispeaker environments, where disentangling and independently manipulating prosodic and linguistic features is essential for high-quality synthesis.

Moreover, encoding features conditioned on predictions from past quantized latent frames enhances temporal coherence, contributing to more stable audio outputs [63]. This technique underscores the importance of temporal dynamics in phoneme-level representations, ensuring that synthesized speech flows naturally.

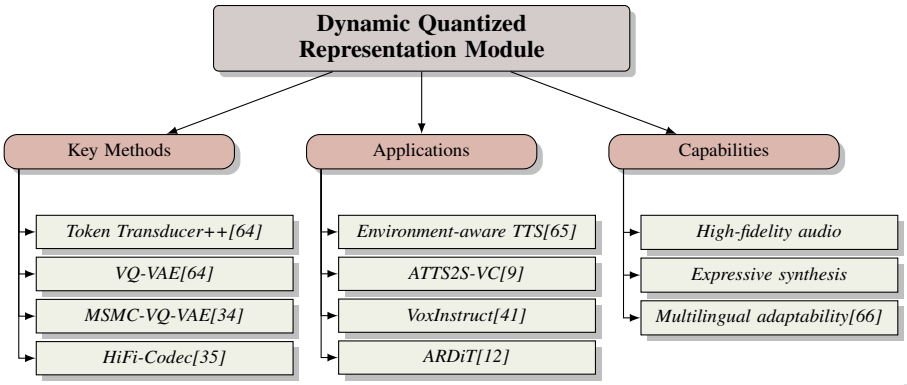


Figure 3: This figure illustrates the hierarchical structure of the Dynamic Quantized Representation Module in TTS systems, categorizing key methods, applications, and capabilities that enhance audio quality and expressiveness.

Phoneme-level discrete latent representations have also been applied innovatively, such as synthesizing high-fidelity piano audio from MIDI input, demonstrating their versatility beyond conventional TTS tasks [68]. This potential opens new avenues for exploration in audio synthesis and related fields.

The application of phoneme-level discrete latent representations significantly contributes to TTS quality by enabling detailed control over prosodic features, enhancing speaker adaptation, and improving the temporal coherence of synthesized speech. As research progresses, the development of sophisticated systems increasingly relies on improved representations critical for enhancing the naturalness, multilingual support, and emotional expression of synthesized speech. This evolution is driven by advancements in algorithms, model pre-training techniques, and high-quality datasets, influencing the performance and accessibility of voice-based technologies across various applications [6, 15, 16, 30].

3.3 WaveNet Autoencoders and Discrete Representation

WaveNet autoencoders significantly advance TTS technologies by learning discrete latent representations that capture phonetic information while discarding speaker-related details [69]. This capability is crucial for generating high-quality, natural-sounding speech, as it allows the synthesis process to focus on linguistic content rather than speaker-specific characteristics. The methodology involves encoding speech into a content code, which is decoded into speech using WaveNet, conditioned on a speaker ID [70]. This approach facilitates the separation of phonetic content from speaker identity, enabling more flexible and adaptable TTS systems.

WaveNet autoencoders enhance the ability to generate speech that is both contextually relevant and acoustically consistent, addressing challenges of traditional TTS methods that struggle to maintain synthesis consistency. By utilizing discrete representation techniques such as Vector Quantized Variational Autoencoders (VQ-VAEs), they effectively capture and synthesize essential phonetic and prosodic elements of speech, preserving the intended nuances of the original audio while enhancing overall naturalness and expressiveness [70, 69].

Moreover, integrating WaveNet autoencoders into TTS architectures supports the development of robust and scalable systems capable of adapting to various linguistic contexts and speaker profiles. This adaptability is particularly advantageous in scenarios involving multiple languages and speakers, where effectively separating and independently adjusting phonetic characteristics and speaker-specific traits is crucial for producing high-quality speech synthesis. Such flexibility enhances the performance of models like Tacotron, which synthesizes speech directly from text and can be trained end-to-end, as well as discrete token-based speech language models that excel in generating varied prosody and spontaneous outputs [49, 39, 28].

As research in speech synthesis progresses, integrating WaveNet autoencoders for extracting discrete representations is anticipated to significantly enhance TTS technologies. This approach aims to improve the quality, efficiency, and versatility of speech synthesis systems by enabling the extraction of phonetic content while minimizing the influence of extraneous factors such as pitch variations and

background noise. Recent advancements, including the use of VQ-VAEs and generative adversarial networks (GANs), have demonstrated promising results in generating high-quality speech waveforms, facilitating more natural and intelligible speech outputs across a wide range of applications, including voice conversion and parametric TTS systems [71, 69, 70, 72].

4 Vector Quantization and Codebook Learning

Category	Feature	Method
Vector-Quantized GAN and Codebook Learning	Quantization Techniques	VR[40], WT[18], LC[21], SC[73], CWT-Vocoder[37], TF-Codec[63], VSASM[66], ATVC[14]
	VQ-GAN and Codebook Techniques	DS[32]
Adversarial Vocoding and Vector Quantization	Adversarial Techniques	AdVoc[74], EATS[75], VI[41], DTC[76]
Extended VQ-VAE and Multi-Encoder Approaches	Multi-Encoder Strategies	EVQ-VAE[77]
Mechanisms of Vector Quantization	Latent Representation	WAE[69], WNLP[72], NL-PVQ[78], PSQ[79], PHEME[80], VQMIVC[56]
Codebook Learning Techniques	Ordered Representation Strategies	OPQ[11]
	Residual and Group Techniques	GRVQ[35]
	Multi-Stage Approaches	MSMC-VQ-VAE[34]

Table 1: This table provides a comprehensive overview of various methods and techniques employed in vector quantization and codebook learning for text-to-speech synthesis. It categorizes the methods into five key areas: Vector-Quantized GAN and Codebook Learning, Adversarial Vocoding and Vector Quantization, Extended VQ-VAE and Multi-Encoder Approaches, Mechanisms of Vector Quantization, and Codebook Learning Techniques, highlighting their respective features and methodologies.

In text-to-speech (TTS) synthesis, vector quantization and codebook learning are critical for encoding speech signals efficiently, enhancing both audio quality and synthesis efficiency. Table 1 summarizes the diverse methodologies and techniques utilized in the integration of vector quantization and codebook learning within text-to-speech synthesis systems. Table 3 offers a comprehensive comparison of various methodologies integrating vector quantization and codebook learning in text-to-speech synthesis, elucidating their impact on speech quality, efficiency, and adaptability. This section explores the integration of vector-quantized generative adversarial networks (VQ-GANs) with codebook learning, highlighting improvements in speech naturalness and control.

4.1 Vector-Quantized GAN and Codebook Learning

Vector-quantized generative adversarial networks (VQ-GANs) combined with codebook learning have significantly enhanced TTS by improving speech quality and control. VQ-GANs encode phonetic information into discrete representations, decoded into high-quality audio, reducing artifacts and discontinuities at phoneme boundaries [18]. GANs enable parallel inference, increasing efficiency over traditional sequential methods [73].

The integration with advanced decoders like HiFi-GAN exemplifies their capability in compressing and enhancing speech quality, providing a robust framework for noise-robust synthesis [76]. The MCRVQ mechanism improves codec generation by distributing information effectively across codebook channels [21], while wavelet-based techniques with Gaussian mixture modeling enhance adaptability [37].

As illustrated in Figure 4, the key components and advancements in Vector-Quantized GAN and Codebook Learning for speech synthesis are highlighted, showcasing enhancements in speech quality, various synthesis techniques, and practical applications. VALL-E R maintains strict alignment between phonemes and audio tokens, enhancing speech quality [40]. The VQ-VAE framework reduces temporal redundancy, improving coding efficiency [63]. This approach provides real-time insights into vowel learning, linked to vector-quantized GANs in TTS [66]. GAN-based adversarial training generates glottal waveforms effectively [13], with multilingual codec language models relating to VQ-GANs and codebook learning [41]. These methods effectively transfer prosody, resulting in natural-sounding speech [14].

The integration of VQ-GANs and codebook learning in TTS systems marks a significant leap in achieving high-quality, efficient, and adaptable speech synthesis. These innovations address real-world speech data variability, leveraging abundant audio sources like YouTube and podcasts to

improve synthesis alignment and intelligibility. Quantized latent representations support practical applications, including voice cloning and speaker anonymization, by optimizing data transfer and minimizing information leakage [33, 50].

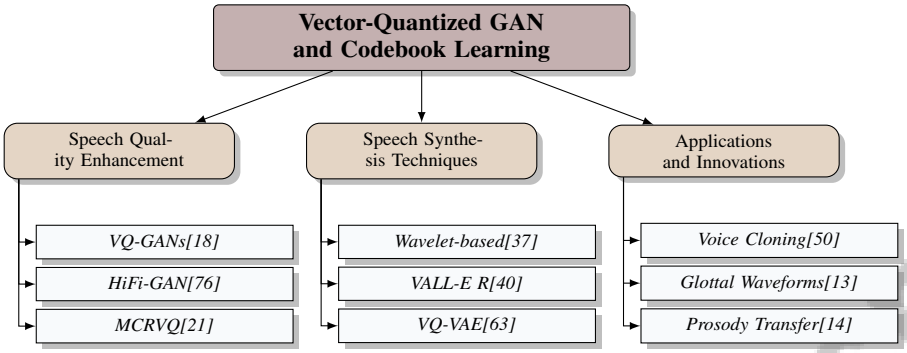


Figure 4: This figure illustrates the key components and advancements in Vector-Quantized GAN and Codebook Learning for speech synthesis, highlighting enhancements in speech quality, various synthesis techniques, and practical applications.

4.2 Adversarial Vocoding and Vector Quantization

Adversarial vocoding techniques, combined with vector quantization, significantly enhance TTS systems by improving the naturalness and quality of synthesized speech. These techniques use generative adversarial networks (GANs) to refine vocoding, generating high-fidelity audio outputs that closely mimic human speech. GANs with vector quantization encode acoustic features into discrete representations, decoded into waveforms with minimal distortion [13].

Adversarial vocoding addresses limitations of traditional methods, which struggle with natural-sounding speech due to sequential processing. GANs enable parallel inference, improving synthesis speed and efficiency [75]. This reduces computational complexity and enhances audio quality by minimizing artifacts and phoneme boundary discontinuities [74].

Integrating adversarial vocoding with vector quantization develops robust TTS systems adaptable to diverse environments and speaker profiles. This adaptability is crucial in multilingual and multispeaker contexts, where disentangling phonetic and speaker features is essential for high-quality synthesis [41]. Adversarial training enhances TTS systems' resilience to noise and input variability, ensuring consistent audio outputs [76].

This integration marks a substantial leap in TTS technology, enhancing synthesis quality, efficiency, and adaptability across diverse datasets. GANs streamline spectrogram conversion, accelerating the synthesis process while maintaining high naturalness and expressiveness. Vector quantization creates discrete latent spaces, optimizing data representation and enhancing speaker anonymization, broadening TTS applicability in real-world scenarios [33, 50, 74, 17].

4.3 Extended VQ-VAE and Multi-Encoder Approaches

Extended VQ-VAE (EVQ-VAE) models advance TTS synthesis by incorporating multi-encoder architectures that capture segmental and suprasegmental features. This dual-encoder approach enhances prosody and naturalness, offering precise control over prosodic elements like intonation, stress, and rhythm [77]. Joint optimization of vocoders with conversion models minimizes data distribution mismatches, improving speech quality and consistency across linguistic contexts [81].

EVQ-VAE models underscore advanced neural architectures' importance in refining synthesis. Multi-encoder architectures enhance processing and generation of high-fidelity audio, crucial for achieving naturalness and expressiveness. Techniques like tuning autoregressive loop granularity, using GANs for acoustic modeling, and integrating VAEs enhance audio quality, bridging the gap between synthesized and human-like speech. Multi-modal context models like MMCE-Qformer improve performance in diverse TTS scenarios, including those requiring longer contextual information [15, 25, 82, 83].

As research evolves, further refinements in EVQ-VAE models and multi-encoder approaches are expected to significantly improve speech synthesis quality, efficiency, and versatility, paving the way for sophisticated and adaptable TTS systems.

4.4 Mechanisms of Vector Quantization

Method Name	Data Representation	Efficiency and Scalability	Integration with Architectures
WAE[69]	Discrete Latent Representations	Reduce Latency	Enhance Synthesis Quality
PHEME[80]	Speech Tokenization	Parallel Processing	Compact Architecture
WNLP[72]	LP Approximation	Parallel Processing	Wavenet Framework
VQMIVC[56]	Content Encoding	Parallel Processing	Neural Network Architectures
NL-PVQ[78]	Vector Quantization	Parallel Processing	Neural Networks
PSQ[79]	Lower-dimensional Space	Real-time Applications	Neural Network

Table 2: Comparison of various vector quantization methods for text-to-speech (TTS) synthesis, highlighting their data representation strategies, efficiency and scalability features, and integration with neural network architectures. The table includes methods such as WAE, PHEME, WNLP, VQMIVC, NL-PVQ, and PSQ, each contributing to advancements in TTS by optimizing latency, processing capabilities, and synthesis quality. This comprehensive overview aids in understanding the role of vector quantization in enhancing TTS systems.

Vector quantization (VQ) is fundamental in neural networks for TTS synthesis, transforming continuous data into discrete representations. This process encodes speech signals into compact forms, facilitating efficient, high-quality audio generation. Autoencoding neural networks learn latent representations of speech signals, focusing on phonetic content essential for capturing speech nuances [69].

VQ optimizes the balance between model complexity and performance. Higy et al.'s benchmark study evaluates metrics for discrete representations, highlighting the importance of designing VQ layers to maximize TTS systems' efficiency and quality [84].

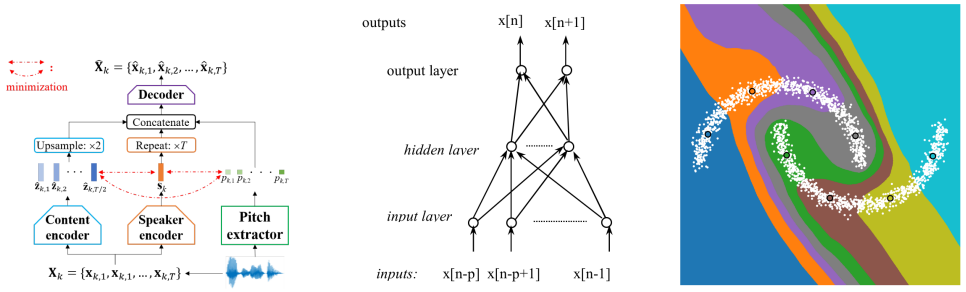
VQ facilitates non-autoregressive decoding, enabling parallel token processing, reducing latency, and improving efficiency, making TTS systems suitable for real-time applications [80]. Parallel processing enhances scalability in handling large-scale speech data.

Integrating VQ with advanced architectures, like LP-WaveNet, improves synthesis quality by modeling excitation generation and LP synthesis filter processes together, reducing mismatch problems and resulting in natural audio outputs [72].

The Wave-U-Net discriminator highlights VQ's potential in maintaining high speech quality with smaller model sizes and faster training times [85]. VQ optimizes neural network architectures for TTS applications, achieving high-quality synthesis with minimal resources.

VQ enhances neural networks for TTS synthesis, advancing efficiency, scalability, and quality. TTS models effectively handle real-world speech's diversity and spontaneity, addressing training-inference alignment mismatches, improving synthesis intelligibility, and enabling voice cloning and speaker anonymization. VQ allows automatic decomposition of prosodic features, enhancing control over speaking styles and outperforming traditional models [33, 86, 50]. Further innovations are expected to refine TTS technologies' capabilities, contributing to AI and machine learning applications in speech synthesis. Table 2 provides a comparative analysis of different vector quantization methods employed in text-to-speech (TTS) synthesis, focusing on their data representation, efficiency, scalability, and integration with various architectures.

As shown in Figure 5, Vector quantization is a fundamental technique in signal processing and data compression, and understanding its mechanisms is crucial for advancing technologies such as speech synthesis and neural network architectures. The provided example illustrates three distinct applications of vector quantization, each highlighting a unique aspect of this technique. The first image showcases a neural network architecture specifically designed for speech synthesis, emphasizing the transformation of input features into speech through a series of encoding processes involving a content encoder, a speaker encoder, and a pitch extractor. The second image presents a more traditional neural network architecture comprising input, hidden, and output layers, demonstrating the flow of information through the network and the interconnections between its nodes. Lastly, the third image captures a complex terrain with colorful contours and points, which can be interpreted as a visual representation of vector quantization in a spatial context, where distinct regions and



(a) A neural network architecture for speech synthesis[56]

(b) A neural network architecture with input, hidden, and output layers[78]

(c) Complex Terrain with Colorful Contours and Points[79]

Figure 5: Examples of Mechanisms of Vector Quantization

patterns are delineated by contours and marked by data points. Together, these examples provide a comprehensive overview of vector quantization's diverse applications and its role in enhancing computational models and data analysis. [56, 78, 79]

4.5 Codebook Learning Techniques

Codebook learning techniques are crucial for advancing TTS systems by providing a framework for efficiently encoding and decoding speech into discrete representations. These techniques optimize the balance between compression efficiency and audio fidelity, crucial for generating high-quality speech [8]. Group Residual Vector Quantization (GRVQ) combines multiple residual vector quantization outputs from grouped latent features, achieving excellent performance with only four codebooks, maintaining high-quality synthesis with fewer resources [35].

The Multi-Stage Multi-Codebook (MSMC) approach enables high-quality synthesis with fewer model parameters, maintaining performance across configurations and ensuring consistent audio quality with reduced computational demands [34]. The SoCodec model introduces Ordered Product Quantization (OPQ), learning an ordered multi-stream representation to improve TTS quality and efficiency, encoding complex speech features into manageable forms [11].

Integrating codebook learning with advanced neural architectures, such as multi-stream representations, underscores these methods' importance in refining TTS systems. Ordered quantization strategies enhance system capability to produce high-fidelity audio, addressing challenges in synthesizing expressive voices and improving TTS technology. Generative Adversarial Networks, Variational Auto-Encoders, and semi-supervised learning approaches generate speech closely matching human recordings in naturalness and intelligibility, even in low-resource scenarios. This progress contributes to TTS systems' evolution, facilitating applications across fields like accessibility, navigation, and virtual assistance [6, 25, 87, 88].

As research evolves, further refinements in codebook learning techniques are expected to drive significant improvements in speech synthesis quality, efficiency, and adaptability. These advancements will enhance TTS systems' functionality, broadening applications across sectors like accessibility—assisting individuals with visual impairments—entertainment, providing lifelike voiceovers, and education, facilitating personalized learning through interactive spoken feedback. Innovations in model architectures, training methodologies, and high-quality datasets enable more natural, multilingual, and emotionally expressive synthesized speech [6, 16, 89, 90].

5 Large Language Models and Neural Codecs

The fusion of large language models (LLMs) and neural codecs has significantly advanced text-to-speech (TTS) synthesis, improving both audio quality and efficiency. This integration leverages LLMs' linguistic strengths and the sophisticated architectures of neural codecs. The following subsections explore the mechanisms and innovations driving this synergy, highlighting its impact on TTS systems.

Feature	Vector-Quantized GAN and Codebook Learning	Adversarial Vocoding and Vector Quantization	Extended VQ-VAE and Multi-Encoder Approaches
Speech Quality	High, Natural, Controlled	Natural, High-fidelity	Enhanced Prosody, Naturalness
Efficiency	Parallel Inference Enabled	Parallel, Reduced Complexity	Joint Optimization, Minimal Mismatch
Adaptability	Noise-robust Synthesis	Multilingual, Multispeaker	Multi-modal, Diverse Scenarios

Table 3: This table provides a comparative analysis of three advanced methodologies in text-to-speech synthesis: Vector-Quantized GAN and Codebook Learning, Adversarial Vocoding and Vector Quantization, and Extended VQ-VAE and Multi-Encoder Approaches. It highlights key features such as speech quality, efficiency, and adaptability, offering insights into the strengths and applications of each approach in enhancing naturalness and robustness in synthesized speech.

5.1 Integration of Large Language Models with Neural Codecs

Integrating LLMs with neural codecs has propelled TTS synthesis forward by enhancing audio quality and operational efficiency. This synergy harnesses LLMs' linguistic capabilities alongside neural codecs' robust encoding and decoding features. For example, the SpeechX framework demonstrates improved performance across speech generation tasks by combining the strengths of both LLMs and neural codecs [44]. The multi-scale codec approach, exemplified by MsCodec, optimizes speech compression by capturing diverse information densities of speech features, enhancing synthesis quality with reduced latency, crucial for real-time applications [91].

Hao et al.'s benchmark evaluations highlight methods for augmenting LLMs with TTS capabilities, facilitating strategy comparisons and emphasizing LLMs' potential to enhance TTS [38]. The Fish-Speech integration strategy exemplifies LLMs' advantages by enabling direct linguistic feature extraction without explicit grapheme-to-phoneme conversion, streamlining synthesis [38]. The TF-Codec achieves superior low-latency speech coding, delivering high-quality audio at lower bitrates than traditional codecs like Opus and EVS, optimizing TTS for real-time applications [63]. The VoxInstruct model further enhances synthesis quality and efficiency through LLM integration [41].

The GEN model's ability to generate language-specific parameters underscores LLMs' integration with neural codecs, enhancing multilingual synthesis and addressing linguistic diversity challenges. The VC-based Polyglot TTS framework synthesizes cross-lingual speech by employing a voice conversion model to create synthetic datasets for training monolingual acoustic models, showcasing LLMs' versatility in TTS applications [4]. The ARDiT model innovatively generates continuous audio tokens in a single step, significantly reducing latency and enhancing TTS performance [12].

Integrating LLMs with neural codecs enhances TTS systems by improving synthesis quality, reducing data requirements, and increasing adaptability. As research continues to explore synergies between advanced algorithms and linguistic models, further innovations are anticipated, impacting sectors like accessibility, navigation, and virtual assistance, driving the evolution of AI and machine learning in speech synthesis [6, 8].

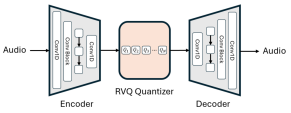
5.2 Innovations in Neural Codec Architectures

Recent advancements in neural codec architectures have significantly influenced TTS systems, enhancing audio synthesis quality and efficiency. Notable innovations include SQ-based quantization techniques, which simplify traditional vector quantization methods, enhancing codec efficiency by reducing computational complexity while maintaining high audio quality [79]. These techniques effectively address high-dimensional data representation challenges, making them suitable for real-time TTS applications.

The development of ESPnet-Codec offers a unified evaluation toolkit, VERSA, supporting a comprehensive range of audio evaluation metrics, allowing thorough codec performance analysis and identifying improvement areas in TTS systems [92]. Using random codebooks in neural codec architectures provides robustness against codebook collapse and reduces training complexity, contributing to stable and efficient TTS systems that maintain high-quality audio synthesis across diverse linguistic and acoustic contexts [93].

The HiFi-Codec exemplifies advanced neural codec architecture potential by achieving high-quality audio reconstruction with fewer codebooks than existing methods. This efficiency is attained through group residual vector quantization, optimizing the balance between compression and audio fidelity, ensuring that TTS systems deliver high-quality outputs with minimal computational resources [35].

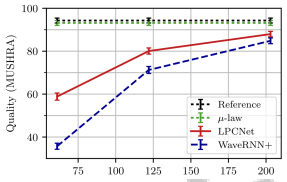
These innovations in neural codec architectures promise to enhance TTS systems' performance and versatility, paving the way for more sophisticated and adaptable speech synthesis technologies. As TTS synthesis research progresses, significant improvements in efficiency, scalability, and audio quality are anticipated, driven by innovative approaches such as integrating high-quality, language-specific datasets, enhancing algorithms for natural-sounding speech, and employing techniques like transfer learning to optimize model performance with limited training data. Furthermore, developing TTS-driven data augmentation methods aims to improve non-autoregressive models' quality, expanding TTS technology applicability across sectors, including accessibility tools, navigation systems, and virtual assistants [6, 16, 7, 42].



(a) Audio Quantization with Convolutional Neural Networks[82]

Layer	Type	# Parameters		GOPs/second audio	
		Per-layer	Total	Per-layer	Total
Pre-processing Layers					
Embedding	Embedding		30,720		
Feature Upsample	ConvTranspose1d		5,120,080		0.82
Repeated Layers					
Dilation	Dilated Conv1d	57,840	925,440	1.84	29.49
Conditional	Conv1d	19,440	311,040	0.61	9.83
Residual	Conv1d	14,520	217,800	0.46	6.91
Skip	Conv1d	29,040	464,640	0.92	14.75
Post-processing Layers					
Out	Conv1d		61,440		1.96
End	Conv1d		65,536		2.09
Total			7,196,696		65.85

(b) The image shows a table with the following information:[57]



(c) Comparison of Quality Metrics for Speech Recognition Systems[94]

Figure 6: Examples of Innovations in Neural Codec Architectures

As illustrated in Figure 6, the exploration of LLMs and neural codecs has ushered in a new era of innovation within neural codec architectures, yielding significant advancements in audio processing and speech recognition technologies. The first example highlights the use of convolutional neural networks (CNNs) in audio quantization, employing an encoder-decoder framework to process audio signals through convolutional layers before quantizing using the Random Vector Quantization (RVQ) algorithm. The second example presents a detailed table outlining the parameters and computational demands of various neural network layers involved in audio processing, showcasing the intricate architecture required for efficient performance. Lastly, the comparison of quality metrics for speech recognition systems provides insights into the effectiveness of different models, such as Reference, -law, LPCNet, and WaveRNN+, emphasizing the importance of training sample size on output quality measured in MUSHRA units. Collectively, these examples underscore the transformative impact of neural codec architectures in enhancing audio and speech processing capabilities [82, 57, 94].

5.3 Case Study: VALL-E and Related Models

The VALL-E model signifies a considerable advancement in TTS synthesis, utilizing neural codecs to enhance speech quality and system efficiency. By employing a conditional language model, VALL-E generates high-fidelity speech from phoneme sequences and acoustic prompts, facilitating diverse and contextually appropriate outputs while accommodating user preferences with zero-shot capabilities. The model's architecture supports speaker-language disentanglement, enhancing TTS synthesis compared to standard polyglot approaches [40]. VALL-E R further improves the model by introducing a phoneme monotonic alignment strategy and codec-merging approach, significantly enhancing inference speed and robustness [40].

Complementary to VALL-E, the MMCE-Qformer model emphasizes advancements in naturalness and speaker similarity in TTS synthesis, effectively leveraging long contextual information to outperform baseline models in various evaluation scenarios. This illustrates the potential of integrating neural codecs with advanced language models for superior synthesis quality. Studies indicate that quantizing mel-spectrograms does not compromise audio fidelity while improving synthesis performance, which is crucial for models like VALL-E [11].

The CLMSTTS method exemplifies the capability to synthesize intelligible and natural speech for unseen speakers across languages, demonstrating the versatility of neural codecs in cross-lingual TTS applications, essential for expanding TTS technologies into diverse linguistic contexts [36]. The SoCodec model, evaluated against baseline systems like X-TTS and VALL-E, focuses on performance metrics such as naturalness and similarity scores, further emphasizing the efficacy of neural codecs in TTS synthesis [11].

Moreover, the HiFi-Codec utilizes group residual vector quantization to achieve high-quality audio reconstruction with fewer codebooks, optimizing the balance between compression and audio fidelity. Future work will concentrate on further optimizing HiFi-Codec and validating its performance across more downstream tasks [35]. The effectiveness of these models underscores the importance of neural codecs in refining TTS systems, enhancing adaptability and performance across applications.

The case study on VALL-E and related models highlights the transformative impact of neural codecs in TTS synthesis, driving advancements in speech quality, efficiency, and adaptability. Innovations in TTS synthesis, including neural network-based models and high-quality dataset generation tools, are set to significantly enhance the naturalness, expressiveness, and efficiency of synthesized speech. By leveraging cutting-edge algorithms and robust data augmentation techniques, these improvements promise to refine TTS capabilities and lay the groundwork for future advancements in artificial intelligence and machine learning across various sectors, including accessibility, navigation, and virtual assistance [16, 8, 95, 6, 39].

6 End-to-End Speech Synthesis

6.1 Concept and Advantages of End-to-End Speech Synthesis

End-to-end speech synthesis represents a transformative approach in text-to-speech (TTS) systems, integrating the entire process from text input to audio output within a single framework. Unlike traditional TTS systems that separate text analysis, acoustic modeling, and vocoding, end-to-end systems streamline these components, enhancing efficiency and speech quality. The ARDiT model exemplifies this by generating high-fidelity audio without discrete tokenization, thus improving reconstruction accuracy and reducing latency [12].

These systems leverage extended contextual information to produce more natural and expressive speech. The VALL-E R model enhances zero-shot TTS synthesis through phoneme prediction and monotonic alignment strategies [4], while the SoCodec model compresses speech sequences effectively, outperforming traditional methods [11].

End-to-end TTS systems also improve training efficiency and reduce data requirements by eliminating the need for time-aligned data, enabling faster training and broader language applicability [3]. Innovations like FLY-TTS, which employs grouped parameter-sharing and a rapid decoder architecture, demonstrate potential reductions in inference time and model size while maintaining high-quality audio [1]. Advanced neural architectures, such as CWT-based vocoders, further enhance the naturalness of synthesized speech compared to state-of-the-art methods like WaveNet [3].

End-to-end speech synthesis thus offers significant advantages over traditional systems, including enhanced naturalness, expressiveness, and efficiency. These advancements are poised to refine TTS capabilities, contributing to the evolution of AI and machine learning applications, especially as future research optimizes diffusion models for real-time applications and improves audio generation efficiency [96].

6.2 Performance Insights and Applications

Benchmark	Size	Domain	Task Format	Metric
EnCodec[97]	1,736,000	Speaker Recognition	Speaker Verification	EER, DER
Bark[49]	4,800	Speech Synthesis	Text-to-Speech	Word Error Rate, Mean Opinion Score
MOSNet[98]	3,200	Speech Quality Assessment	Quality Estimation	SRCC, KTAU
IndicTTS[90]	272,000	Speech Synthesis	Text-to-Speech	MOS, MCD
TTS-ASR[99]	28,000	Automatic Speech Recognition	Synthetic Data Generation	WER, NISQA MOS
LibriSpeech-100h[100]	100,000	Speech Recognition	Word Error Rate (wer)	WER
MSTTS[101]	199,257	Speech Synthesis	Text-to-Speech	Mean Opinion Score, Differential MOS
RobustVoc[102]	38,139	Speech Synthesis	Speech Generation	Mean Opinion Score

Table 4: This table presents a comprehensive overview of various benchmarks used in the evaluation of speech synthesis and recognition systems. It details the size, domain, task format, and performance metrics associated with each benchmark, providing a valuable resource for assessing the capabilities and effectiveness of different models in the field of speech technology.

End-to-end speech synthesis systems have made substantial progress in performance metrics and practical applications, significantly enhancing the naturalness and expressiveness of synthesized speech. The integration of deep learning techniques has been pivotal, transitioning from traditional models to more robust systems. The NNSS method, for instance, produces more natural-sounding speech compared to conventional systems, underscoring its performance benefits [103]. The R-MelNet model, despite slower sampling speeds, delivers expressive audio outputs with fewer parameters, highlighting its efficiency [104].

Performance evaluations using metrics such as mean opinion score (MOS) and objective assessments like PESQ confirm these systems' ability to deliver high-quality audio. DelightfulTTS 2 achieved a CMOS gain of +0.14 over its predecessor, demonstrating its effectiveness in producing superior audio quality [17]. Additionally, models utilizing EnCodec tokens perform competitively with those using mel-spectrogram features, often within 1

Innovative models like ZMM-TTS exhibit high naturalness and speaker similarity in synthesized speech, even for unseen speakers across multiple languages, demonstrating adaptability in multilingual contexts [23]. The Wave-U-Net discriminator replaces an ensemble of discriminators while maintaining comparable speech quality, significantly reducing model size and enhancing processing speed [85], which is crucial for real-time applications with constrained computational resources.

Additionally, Fish-Speech outperforms baseline models in managing complex linguistic scenarios and voice cloning tasks, highlighting the potential of large language models to enhance TTS systems [38]. The optimization of neural speech codecs is validated by experimental results focusing on metrics that quantify audio quality [91].

Advancements in end-to-end speech synthesis systems offer significant benefits in performance and practical applications, paving the way for more sophisticated and adaptable TTS technologies. These innovations enhance TTS deployment across diverse fields, including accessibility, entertainment, and real-time communication, while addressing challenges such as computational demands and the need for extensive datasets [105]. As research continues, further refinements are anticipated to propel TTS technology forward, contributing to the broader development of AI and machine learning applications in speech synthesis.

7 Applications and Future Directions

7.1 Current Applications of Advanced TTS Technologies

Advanced text-to-speech (TTS) technologies have revolutionized various sectors by enhancing communication, accessibility, and personalization. In multilingual speech synthesis, these technologies adeptly handle linguistic complexities, crucial for global communication platforms [2]. On-device TTS systems bolster user privacy by minimizing cloud dependency, ensuring secure data management while delivering high-quality speech synthesis, vital for real-time applications like virtual assistants [95]. Systems such as Voicebox and Vec-Tok Speech exemplify TTS versatility, enabling zero-shot TTS, content editing, voice conversion, and speaking style transfer for personalized outputs [106, 107].

In accessibility, TTS systems offer inclusive communication tools tailored to specific accents and speech patterns, supporting individuals with speech impairments [108]. Advances in accent conversion further enhance inclusivity [109]. TTS technologies also generate synthetic data to improve automatic speech recognition (ASR) systems, especially in low-resource settings, enhancing ASR accuracy and reliability [100]. Voice cloning technologies like NAUTILUS and Make-A-Voice facilitate realistic voice replicas and scalable, controllable voice synthesis, impacting entertainment and personalized communication [110, 111].

These advancements, driven by deep learning and neural networks, not only produce more natural and expressive speech but also augment data for ASR systems, laying the foundation for future innovations [8, 7, 6, 30, 39].

7.2 Enhancements in Model Efficiency and Scalability

Recent TTS advancements focus on enhancing efficiency and scalability to accommodate diverse linguistic contexts. Expanding TTS models to additional languages significantly boosts scalability in

multilingual applications [112]. Multimodal speech processing and self-supervised learning reduce the need for extensive labeled datasets, enhancing model efficiency [29]. Transfer learning facilitates efficient adaptation to diverse speaker characteristics, personalizing outputs and streamlining training [113].

Improving video-to-speech (VTS) systems for cross-domain and multilingual scenarios underscores TTS scalability importance [114]. Generative language modeling techniques applied to discrete speech tokens enable scalable, context-aware speech language models (SLMs) to produce diverse, natural outputs, excelling in prosody and spontaneous behavior generation. As TTS systems evolve, they are expected to deliver high-quality, personalized synthesis across applications, including accessibility tools and virtual assistants. Addressing intelligibility and speaker consistency challenges remains crucial [49, 6].

7.3 Adaptability to Diverse Languages and Accents

TTS adaptability to diverse languages and accents is crucial for personalized, contextually relevant synthesis. Recent advancements enhance linguistic diversity, enabling accurate replication across languages and accents [2]. Multilingual models leverage shared features for improved synthesis quality, employing language-agnostic acoustic modeling and speaker adaptation [4]. Zero-shot learning allows synthesis for languages and accents absent in training data, generalizing across linguistic boundaries [23].

Accent conversion methodologies enhance adaptability, enabling seamless conversion between accents [109]. These innovations promise to improve TTS accessibility and effectiveness across applications, enhancing naturalness, multilingual capabilities, and emotional expressiveness [16, 8, 95, 6, 39].

7.4 Integration with Emerging Technologies

Integrating TTS with deep learning and dataset generation advancements enhances synthesized speech's naturalness and accessibility, enabling innovative applications in accessibility tools and virtual assistants [6, 16, 90]. TTS integration with VR/AR platforms creates realistic, interactive experiences by converting text into natural, contextually relevant audio. These systems enhance communication by adapting to acoustic environments and incorporating emotional expression, improving accessibility in navigation aids and virtual assistants [65, 115, 7, 6, 116].

TTS integration with smart home devices enhances interaction through natural communication, vital for accessibility and navigation. On-device processing and continuous speech tokenization pave the way for efficient, high-quality voice outputs [16, 117, 95, 108, 6]. TTS technology integration with edge computing and IoT promises real-time, high-quality voice output, enhancing user experiences in accessibility tools and virtual assistants [6, 95]. Processing speech data locally on edge devices delivers real-time audio outputs while minimizing cloud reliance, enhancing data privacy and security.

In healthcare, TTS facilitates natural communication, enhancing accessibility for patients with speech or reading difficulties and supporting healthcare professionals in improving interactions. Recent innovations using deep convolutional neural networks make TTS systems more efficient and cost-effective [6, 30]. TTS systems for assistive technologies provide personalized communication tools, integrating with telemedicine platforms and wearable devices to improve patient care and accessibility.

TTS integration with emerging technologies holds tremendous promise for advancing capabilities and applications. As TTS synthesis research progresses, significant innovations are anticipated, enhancing accessibility tools, navigation systems, and virtual assistants by converting text into natural, intelligible speech using advanced algorithms and linguistic models [6, 39, 8].

7.5 Future Research Directions in TTS Technologies

The future of TTS technologies focuses on improving robustness, efficiency, and adaptability. Optimizing codec designs and exploring diverse applications lead to more adaptable systems [21]. Enhancements in T2S components and parallel decoding strategies are crucial, focusing on high-quality conversational TTS data. Efforts to enhance expressiveness by making utterance embedding

space interpretable and controllable are vital, allowing manual prosodic pattern selection [13]. Future work should explore multilingual support and vocoder training [9], assessing vocoder impacts on perceptions and improving multilingual synthesis [36].

Refining alignment techniques for longer utterances and exploring additional training data sources are critical [24]. Robustness evaluations of neural vocoders should consider comprehensive assessments [41]. Research should focus on Transformer architectures versus RNNs, extending pretraining methods for nonparallel conditions, enhancing prosody control, reducing labeled data reliance, and exploring expressive synthesis trends [38]. Supporting multiple languages and speech-generation tasks is promising [32].

Future research could optimize model architecture and training processes for diverse languages and dialects [118]. Decoding efficiency optimization and VQ-based lattices exploration for tasks like translation and keyword spotting are areas for exploration. End-to-end architectures for speech-to-text and speech-to-speech translation could influence future directions [66]. Evaluating TTS using learned representations and exploring F0 and waveform feature disentanglement are important directions.

Attention constraints for Transformer-NMT models and multi-speaker scenarios are potential advancements [10]. Research should prioritize improving model efficiency, exploring unsupervised learning, and addressing noise reduction challenges [105]. Enhancing compression-based audio token performance and investigating tokenization methods are future directions [97].

These research directions underline efforts to refine TTS capabilities, contributing to AI and machine learning advancements in speech synthesis. As TTS research progresses, innovations are expected to enhance sophistication and adaptability, meeting diverse needs across applications, including accessibility tools and multilingual support. Key areas include improving naturalness, emotional expression, and integrating high-quality datasets for performance optimization. Exploring neural network architectures and pre-training methodologies promises to refine TTS capabilities for underrepresented languages, fostering inclusivity [15, 16, 8, 6, 90]. Future research could expand datasets and refine evaluation techniques to enhance synthesis quality, while exploring Wave-U-Net discriminator applications in singing and emotional synthesis offers promising avenues.

8 Conclusion

The evolution of text-to-speech (TTS) synthesis has been significantly driven by advancements in discrete representation, vector quantization, and neural codecs, which collectively enhance the quality and naturalness of speech generation systems. Technologies like SoCodec have notably improved the fidelity of synthesized speech, underscoring the importance of ordered product quantization in advancing TTS capabilities. The integration of discrete representations has led to enhanced phonetic precision and speaker independence, facilitating improved generalization across diverse applications.

Neural codecs, particularly spectral codecs, have enabled non-autoregressive models to achieve superior audio quality compared to traditional methods. Innovations such as continuous noise masking in vocoders offer promising solutions to reduce unwanted noise in speech signals. Moreover, models like MQTTS have demonstrated exceptional performance in synthesizing speech from real-world data, achieving higher intelligibility and naturalness.

Future research directions may focus on refining model architectures and training methodologies to boost the efficiency and adaptability of systems like MSMC-TTS. The potential of randomization in quantizers to enhance audio representation and generalization is also worth exploring. Additionally, vowel space analysis could contribute to improving synthetic voice quality, further advancing TTS technology.

Diffusion models present a promising avenue for enhancing audio quality and generation efficiency, although challenges such as the demand for extensive training data and the risk of producing robotic-sounding speech persist. Continued research in discrete representation, vector quantization, and neural codecs is expected to drive substantial progress in TTS technologies. These innovations are poised to create more sophisticated and versatile systems, influencing the future landscape of AI and machine learning applications in speech synthesis. Efforts to simplify dilated convolution blocks and revisit classical speech modeling techniques aim to further enhance model performance, with multilingual TTS advancements holding significant potential for future AI developments.

References

[1] 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, 2024.

[2] Tomáš Nekvinda and Ondřej Dušek. One model, many languages: Meta-learning for multilingual text-to-speech, 2020.

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

[4] Dariusz Piotrowski, Renard Korzeniowski, Alessio Falai, Sebastian Cygert, Kamil Pokora, Georgi Tinchev, Ziyao Zhang, and Kayoko Yanagisawa. Cross-lingual knowledge distillation via flow-based voice conversion for robust polyglot text-to-speech, 2023.

[5] 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.

[6] Harini s and Manoj G M. Text to speech synthesis, 2024.

[7] Min-Jae Hwang, Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis, 2020.

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

[9] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms, 2018.

[10] Tomoki Hayashi and Shinji Watanabe. Discretalk: Text-to-speech as a machine translation problem, 2020.

[11] Haohan Guo, Fenglong Xie, Kun Xie, Dongchao Yang, Dake Guo, Xixin Wu, and Helen Meng. Socodec: A semantic-ordered multi-stream speech codec for efficient language model based text-to-speech synthesis, 2024.

[12] Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. Autoregressive diffusion transformer for text-to-speech synthesis, 2024.

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

[14] 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.

[15] Guangyan Zhang, Yichong Leng, Daxin Tan, Ying Qin, Kaitao Song, Xu Tan, Sheng Zhao, and Tan Lee. A study on the efficacy of model pre-training in developing neural text-to-speech system, 2021.

[16] Ahmet Gunduz, Kamer Ali Yuksel, Kareem Darwish, Golara Javadi, Fabio Minazzi, Nicola Sobieski, and Sebastien Bratieres. An automated end-to-end open-source software for high-quality text-to-speech dataset generation, 2024.

[17] 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.

[18] 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.

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

[20] Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. Wavthruvec: Latent speech representation as intermediate features for neural speech synthesis, 2022.

[21] Shengpeng Ji, Minghui Fang, Ziyue Jiang, Siqi Zheng, Qian Chen, Rongjie Huang, Jialung Zuo, Shulei Wang, and Zhou Zhao. Language-codec: Reducing the gaps between discrete codec representation and speech language models, 2024.

[22] 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.

[23] Cheng Gong, Xin Wang, Erica Cooper, Dan Wells, Longbiao Wang, Jianwu Dang, Korin Richmond, and Junichi Yamagishi. Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations, 2024.

[24] Hui Lu, Zhiyong Wu, Xixin Wu, Xu Li, Shiyin Kang, Xunying Liu, and Helen Meng. Vaenar-tts: Variational auto-encoder based non-autoregressive text-to-speech synthesis, 2021.

[25] 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.

[26] Hongkun Hao, Long Zhou, Shujie Liu, Jinyu Li, Shujie Hu, Rui Wang, and Furu Wei. Boosting large language model for speech synthesis: An empirical study, 2023.

[27] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[28] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

[29] Ambuj Mehrish, Navonil Majumder, Rishabh Bhardwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing, 2023.

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

[31] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining, 2019.

[32] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation, 2023.

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

[34] Haohan Guo, Fenglong Xie, Frank K. Soong, Xixin Wu, and Helen Meng. A multi-stage multi-codebook vq-vae approach to high-performance neural tts, 2022.

[35] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec, 2023.

[36] Jing Xu, Daxin Tan, Jiaqi Wang, and Xiao Chen. Enhancing multilingual speech generation and recognition abilities in llms with constructed code-switched data, 2024.

[37] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Csaba Zainkó, and Géza Németh. Continuous wavelet vocoder-based decomposition of parametric speech waveform synthesis, 2021.

[38] 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.

[39] Zhuangqun Huang, Gil Keren, Ziran Jiang, Shashank Jain, David Goss-Grubbs, Nelson Cheng, Farnaz Abtahi, Duc Le, David Zhang, Antony D'Avirro, Ethan Campbell-Taylor, Jessie Salas, Irina-Elena Veliche, and Xi Chen. Text generation with speech synthesis for asr data augmentation, 2023.

[40] 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.

[41] Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling, 2024.

[42] Ze Liu. Comparative analysis of transfer learning in deep learning text-to-speech models on a few-shot, low-resource, customized dataset, 2023.

[43] Ryuichi Yamamoto, Eunwoo Song, Min-Jae Hwang, and Jae-Min Kim. Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators, 2021.

[44] 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.

[45] Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning, 2022.

[46] Da-Yi Wu, Yen-Hao Chen, and Hung-Yi Lee. Vqvc+: One-shot voice conversion by vector quantization and u-net architecture, 2020.

[47] Zhikang Niu, Sanyuan Chen, Long Zhou, Ziyang Ma, Xie Chen, and Shujie Liu. Ndvq: Robust neural audio codec with normal distribution-based vector quantization, 2024.

[48] Viet Anh Trinh, Rosy Southwell, Yiwen Guan, Xinlu He, Zhiyong Wang, and Jacob Whitehill. Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing, 2024.

[49] Siyang Wang and Éva Székely. Evaluating text-to-speech synthesis from a large discrete token-based speech language model, 2024.

[50] Hieu-Thi Luong and Junichi Yamagishi. Preliminary study on using vector quantization latent spaces for tts/vc systems with consistent performance, 2021.

[51] Joun Yeop Lee, Myeonghun Jeong, Minchan Kim, Ji-Hyun Lee, Hoon-Young Cho, and Nam Soo Kim. High fidelity text-to-speech via discrete tokens using token transducer and group masked language model, 2024.

[52] Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, Zhikang Niu, Shuai Wang, Hui Zhang, Xie Chen, and Kai Yu. Vall-t: Decoder-only generative transducer for robust and decoding-controllable text-to-speech, 2024.

[53] Vincent Wan, Chun an Chan, Tom Kenter, Jakub Vit, and Rob Clark. Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network, 2019.

[54] Shun Lei, Yixuan Zhou, Liyang Chen, Dan Luo, Zhiyong Wu, Xixin Wu, Shiyin Kang, Tao Jiang, Yahui Zhou, Yuxing Han, and Helen Meng. Improving language model-based zero-shot text-to-speech synthesis with multi-scale acoustic prompts, 2024.

[55] Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech, 2019.

[56] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion, 2021.

[57] Sam Davis, Giuseppe Coccia, Sam Gooch, and Julian Mack. Empirical evaluation of deep learning model compression techniques on the wavenet vocoder, 2020.

[58] Edresson Casanova, Ryan Langman, Paarth Neekhara, Shehzeen Hussain, Jason Li, Subhankar Ghosh, Ante Jukić, and Sang gil Lee. Low frame-rate speech codec: a codec designed for fast high-quality speech llm training and inference, 2024.

[59] Yuzhe Gu and Enmao Diao. Esc: Efficient speech coding with cross-scale residual vector quantized transformers, 2024.

[60] Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model, 2024.

[61] Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. Single-codec: Single-codebook speech codec towards high-performance speech generation, 2024.

[62] Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation, 2024.

[63] Xue Jiang, Xiulian Peng, Huaying Xue, Yuan Zhang, and Yan Lu. Latent-domain predictive neural speech coding, 2023.

[64] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[65] Daxin Tan, Guangyan Zhang, and Tan Lee. Environment aware text-to-speech synthesis, 2022.

[66] Binu Abeysinghe, Jesin James, Catherine I. Watson, and Felix Marattukalam. Visualising model training via vowel space for text-to-speech systems, 2022.

[67] Sotirios Karapiperis, Nikolaos Ellinas, Alexandra Vioni, Junkwang Oh, Gunu Jho, Inchul Hwang, and Spyros Raptis. Investigating disentanglement in a phoneme-level speech codec for prosody modeling, 2024.

[68] Xuan Shi, Erica Cooper, Xin Wang, Junichi Yamagishi, and Shrikanth Narayanan. Can knowledge of end-to-end text-to-speech models improve neural midi-to-audio synthesis systems?, 2023.

[69] Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders, 2019.

[70] Mingjie Chen and Thomas Hain. Unsupervised acoustic unit representation learning for voice conversion using wavenet auto-encoders, 2020.

[71] 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.

[72] Min-Jae Hwang, Frank Soong, Eunwoo Song, Xi Wang, Hyeonjoo Kang, and Hong-Goo Kang. Lp-wavenet: Linear prediction-based wavenet speech synthesis, 2020.

[73] Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Boris Ginsburg. Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis, 2024.

[74] Paarth Neekhara, Chris Donahue, Miller Puckette, Shlomo Dubnov, and Julian McAuley. Expediting tts synthesis with adversarial vocoding, 2019.

[75] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech, 2021.

[76] Xue Jiang, Xiulian Peng, Yuan Zhang, and Yan Lu. Disentangled feature learning for real-time neural speech coding, 2023.

[77] Yi Zhao, Haoyu Li, Cheng-I Lai, Jennifer Williams, Erica Cooper, and Junichi Yamagishi. Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction, 2020.

[78] Marcos Faundez-Zanuy. Non-linear predictive vector quantization of speech, 2022.

[79] Andreas Brendel, Nicola Pia, Kishan Gupta, Lyonel Behringer, Guillaume Fuchs, and Markus Multrus. Neural speech coding for real-time communications using constant bitrate scalar quantization, 2024.

[80] Paweł Budzianowski, Taras Sereda, Tomasz Cichy, and Ivan Vulić. Pheme: Efficient and conversational speech generation, 2024.

[81] Haitong Zhang. The neteasegames system for voice conversion challenge 2020 with vector-quantization variational autoencoder and wavenet, 2020.

[82] Jiaqi Li, Dongmei Wang, Xiaofei Wang, Yao Qian, Long Zhou, Shujie Liu, Midia Yousefi, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Yanqing Liu, Junkun Chen, Sheng Zhao, Jinyu Li, Zhizheng Wu, and Michael Zeng. Investigating neural audio codecs for speech language model-based speech generation, 2024.

[83] Jinlong Xue, Yayue Deng, Yicheng Han, Yingming Gao, and Ya Li. Improving audio codec-based zero-shot text-to-speech synthesis with multi-modal context and large language model, 2024.

[84] Bertrand Higy, Lieke Gelderloos, Afra Alishahi, and Grzegorz Chrupała. Discrete representations in neural models of spoken language, 2021.

[85] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Shogo Seki. Wave-u-net discriminator: Fast and lightweight discriminator for generative adversarial network-based speech synthesis, 2023.

[86] Yutian Wang, Yuankun Xie, Kun Zhao, Hui Wang, and Qin Zhang. Unsupervised quantized prosody representation for controllable speech synthesis, 2022.

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

[88] Haohan Guo, Fenglong Xie, Jiawen Kang, Yujia Xiao, Xixin Wu, and Helen Meng. Qs-tts: Towards semi-supervised text-to-speech synthesis via vector-quantized self-supervised speech representation learning, 2023.

[89] Ruiqing Xue, Yanqing Liu, Lei He, Xu Tan, Linquan Liu, Edward Lin, and Sheng Zhao. Foundationtts: Text-to-speech for asr customization with generative language model, 2023.

[90] Gokul Karthik Kumar, Praveen S V au2, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. Towards building text-to-speech systems for the next billion users, 2023.

[91] Peiji Yang, Fengping Wang, Yicheng Zhong, Huawei Wei, and Zhisheng Wang. Optimizing neural speech codec for low-bitrate compression via multi-scale encoding, 2024.

[92] Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, et al. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 562–569. IEEE, 2024.

[93] Benoît Giniès, Xiaoyu Bie, Olivier Fercoq, and Gaël Richard. Using random codebooks for audio neural autoencoders, 2024.

[94] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction, 2019.

[95] Sivanand Achanta, Albert Antony, Ladan Golipour, Jiangchuan Li, Tuomo Raitio, Ramya Rasipuram, Francesco Rossi, Jennifer Shi, Jaimin Upadhyay, David Winarsky, and Hepeng Zhang. On-device neural speech synthesis, 2021.

[96] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai, 2023.

[97] Krishna C. Puvvada, Nithin Rao Koluguri, Kunal Dhawan, Jagadeesh Balam, and Boris Ginsburg. Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition, 2023.

[98] 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.

[99] Nick Rossenbach, Ralf Schlüter, and Sakriani Sakti. On the problem of text-to-speech model selection for synthetic data generation in automatic speech recognition, 2024.

[100] Nick Rossenbach, Mohammad Zeineldeen, Benedikt Hilmes, Ralf Schlüter, and Hermann Ney. Comparing the benefit of synthetic training data for various automatic speech recognition architectures, 2021.

[101] Erica Cooper, Xin Wang, Yi Zhao, Yusuke Yasuda, and Junichi Yamagishi. Pretraining strategies, waveform model choice, and acoustic configurations for multi-speaker end-to-end speech synthesis, 2020.

[102] Po chun Hsu, Chun hsuan Wang, Andy T. Liu, and Hung yi Lee. Towards robust neural vocoding for speech generation: A survey, 2020.

[103] Orhan Karaali, Gerald Corrigan, and Ira Gerson. Speech synthesis with neural networks, 1998.

[104] Kyle Kastner and Aaron Courville. R-melnet: Reduced mel-spectral modeling for neural tts, 2022.

[105] Zhaofeng Shi. A survey on audio synthesis and audio-visual multimodal processing, 2021.

[106] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023.

[107] 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, 2023.

[108] Vinotha R, Hepsiba D, L. D. Vijay Anand, and Deepak John Reji. Empowering communication: Speech technology for indian and western accents through ai-powered speech synthesis, 2024.

[109] Jan Melechovsky, Ambuj Mehrish, Berrak Sisman, and Dorien Herremans. Accent conversion in text-to-speech using multi-level vae and adversarial training, 2024.

[110] Hieu-Thi Luong and Junichi Yamagishi. Nautilus: a versatile voice cloning system, 2020.

[111] 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.

[112] Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion, 2020.

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

[114] Disong Wang, Shan Yang, Dan Su, Xunying Liu, Dong Yu, and Helen Meng. Vcvts: Multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion, 2022.

[115] Efthymios Georgiou and Athanasios Katsamanis. Audiovisual speech synthesis: A brief literature review, 2021.

[116] Yoshifumi Nakano, Takaaki Saeki, Shinnosuke Takamichi, Katsuhito Sudoh, and Hiroshi Saruwatari. vtts: visual-text to speech, 2022.

[117] Yixing Li, Ruobing Xie, Xingwu Sun, Yu Cheng, and Zhanhui Kang. Continuous speech tokenizer in text to speech, 2024.

[118] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers, 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.