Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Speech Synthesis and Personalization under Unimodal and Multimodal Conditions

Wanlin Gao*

Department of Computer Science and Technology, Chongqing Jiaotong University, Chongqing, China

*Corresponding author: wanlingao@icloud.com

Abstract. Recently, there have been notable advancements in TTS technology, with researchers optimizing the efficiency, quality, and flexibility of speech generation through various models. This paper systematically explores end-to-end TTS models based on waveform generation, including Parallel WaveGAN, NaturalSpeech, and Multi-Band MelGAN, each of which has unique features in enhancing real-time generation capabilities and sound quality. Additionally, the paper discusses the development of speech separation and synthesis technologies, highlighting the applications of models like CONTENTVEC in pitch adjustment and speaker information disentanglement. In terms of multimodal technology, speech-to-gesture generation has also seen important breakthroughs, utilizing multimodal information to generate natural gestures. The paper provides a detailed summary of the main datasets used in related research, such as LibriTTS, LJSpeech, and VCTK, aiming to offer reference and guidance for future research in speech generation. Although these technologies have achieved significant advancements in efficiency and multifunctionality, the associated models remain complex and require substantial computational resources, limiting their widespread application in practical scenarios.

Keywords: TTS; SpeechSynthesis; Speech-to-Gesture; Personalized Speech.

1. Introduction

This paper offers a thorough examination of the most recent applications of multimodal and monomodal technologies in the fields of speech generation and gesture synthesis, with a deep dive into cutting-edge research in text-to-speech technology. The paper is divided into several sections, each covering key areas and technological advancements.

The paper first analyzes sophisticated models and explores how end-to-end TTS models generate speech directly from waveforms. This module first introduces several methods for generating waveforms from text, including Parallel WaveGAN [1], NaturalSpeech [2], and Multi-Band MelGAN [3]. It then discusses other speech synthesis methods, such as speech resynthesis using adversarial networks and discrete disentangled self-supervised representations [4-5], Neural Source [6], FastPitch [7], and FastDiff [8], and highlights their importance for speech recognition [9].These models streamline the entire process by reducing intermediate steps, thereby enhancing system efficiency while ensuring high-quality output and minimizing computational complexity [10].

Next, the paper focuses on recent advancements in speech separation and synthesis technologies, particularly the breakthroughs in achieving flexible control and high-quality generation of speech. For example, the CONTENTVEC model significantly enhances the flexibility of speech synthesis systems by effectively disentangling speaker information and adjusting pitch [11]. This allows for the generation of speech that is precise in content and rich in style and emotion. The paper also explores how various technologies achieve flexible control over speech without significant content loss, ensuring that the generated speech maintains high quality across diverse and complex scenarios.

In terms of multimodal applications, the paper delves into innovative uses of speech-to-gesture generation [12], demonstrating how speech signals can be used to generate natural matching gestures. By incorporating advanced multimodal contextual information and adversarial training mechanisms,

researchers can generate more diverse and natural gestures. These technologies not only show outstanding performance in academic research but also present substantial potential in practical applications, such as virtual assistants and human-computer interaction.

Finally, the paper summarizes the important datasets used in these studies, including LibriTTS, LJSpeech, and VCTK, analyzing their crucial roles in model training and performance evaluation. By systematically reviewing the characteristics and applications of these datasets, the paper provides valuable references for future research. In addition to summarizing the current state of development, the paper also addresses the main challenges faced in ongoing research, such as how to further improve model efficiency, stability, and generation quality, and how to maintain superior performance in complex and variable application environments.

2. Monomodal

2.1. Text-to-Speech Generation through Waveforms

2.1.1. Text and waveform

End-to-end Text-to-Speech (TTS) models have become a significant research direction, aiming to generate speech waveforms directly from text without intermediate steps. These models simplify the training process, enhancing overall system efficiency. Researchers are also exploring ways to maintain high-quality outputs while reducing computational complexity.

Parallel WaveGAN is primarily based on Generative Adversarial Networks (GANs) to achieve rapid model generation. It employs a non-autoregressive WaveNet as the generator and is trained through a joint optimization of adversarial loss and the short-time Fourier transform loss at several resolutions, effectively capturing the time-frequency distribution features of real speech waveforms [1]. Figure 1 below illustrates its adversarial training framework.

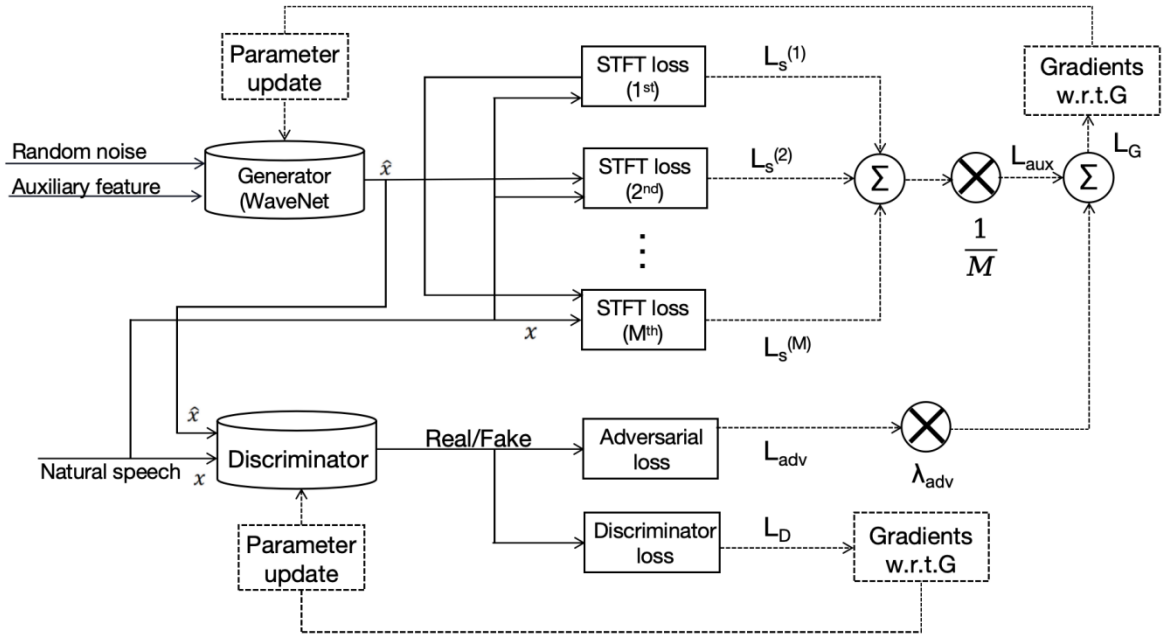

Fig.1 A depiction of its multi-resolution STFT loss-based adversarial training architecture [1]

With its smaller model parameters, it significantly improves the real-time generation of 24kHz speech waveforms on a GPU, achieving speeds 28.68 times faster than real-time [1]. However, this requires substantial computational resources, including high-performance GPUs for training and inference. As a GAN model, it is less stable in training compared to autoregressive models and requires careful parameter tuning to avoid mode collapse, which can result in the absence of diversity within the produced samples.

NaturalSpeech has largely addressed the issues of model stability and waveform quality found in Parallel WaveGAN. As a comprehensive text-to-speech system, NaturalSpeech leverages Variational Autoencoders (VAE) for text-to-waveform generation. It integrates phoneme pre-training, bidirectional prior, and differentiable duration modeling and posterior modeling, and memory-based VAE techniques, leading to a substantial enhancement in model performance [2]. In waveform generation, phoneme pre-training and memory-based VAE are notable breakthroughs in recent years. The model uses a pre-trained phoneme encoder to train phoneme sequence masking language models on large-scale text corpora, enhancing the ability to capture subtle language differences [2]. Additionally, the differentiable duration controller, which bridges the length difference between frame-level posterior and phoneme-level prior, generates flexible features and addresses the training-inference mismatch problem caused by using real values during training and predicted values during inference when training vocoders and Mel-spectrogram decoders [2]. Compared to the previously mentioned technologies, VAE contributes more to waveform generation. It compresses high-dimensional speech into continuous representations at the frame level, resulting in more complex phoneme training. Unlike previous methods, it uses posterior probabilities as queries to access a memory bank and reconstructs waveforms using the results of attention mechanisms. This approach simplifies the entire waveform generation process by only using the posterior result z to calculate attention weights within the memory bank [2]. However, the overall text-to-waveform generation architecture remains complex and demands significant computational resources, particularly in terms of memory usage. Additionally, the training process is time-consuming due to the multiple stages involved.

Muti-Band MelGAN provides a good solution for improving waveform generation speed. Previous waveform generation methods typically relied on generating the current sample based on the previous one and heavily depended on modeling long-term audio dependencies. While these methods produce near-perfect waveform samples, their low efficiency significantly limits practical applications. This approach extends the receptive field to twice that of the original MelGAN, substitutes the multi-resolution STFT loss, which is more significant, for the feature matching loss, and integrates pre-trained multi-band MelGAN, enhancing both the speed and quality of speech waveform generation [3].The Multi-band MelGAN Architecture is shown below.

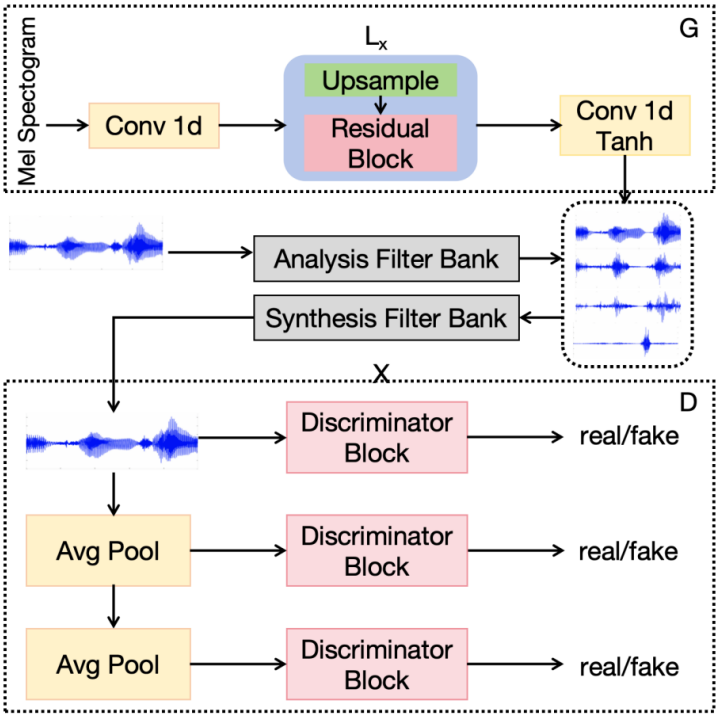

Fig. 2 Multi-band MelGAN Architecture [3]

Figure 2 shows the multi-band MelGAN, an advanced version of the basic MelGAN that adheres to the standard adversarial process between the discriminator and generator. In the MB-MelGAN, the network generator (G) uses the mel spectrogram as a source for producing multiple band indications. Each band's predicted audio signal is first upsampled and then passed through an amalgamation filter. Each band's signals, after being processed by the synthesis filters, are recombined into an audio signal in full band. Subsequently, the discriminator network (D) utilizes the entire band signal as input and, similar to both basic MelGAN and MB-MelGAN, uses multiple discriminators at different scales to distinguish features generated by the generator [3]. Overall, the advantage of Multi-Band MelGAN in real-time synthesis of high-fidelity waveforms has undoubtedly greatly advanced the field of text-to-waveform generation.

2.1.2. Waveforms and Speech

Waveform conversion is another significant area of research in Text-to-Speech (TTS). Researchers have proposed methods to convert speech acoustic features into speech waveforms, including parallel waveform generation techniques based on Generative Adversarial Networks (GANs) [5] and bypassing traditional Mel spectrogram estimation, voice controllable synthesis is achieved by directly using learned speech units in the vocoder [4]. These approaches aim to improve the naturalness and authenticity of generated speech. The application scope of TTS technology is also expanding. For example, TTS can be used in order to enhance the effectiveness of speech recognition systems and to develop methods for assessing the authenticity of audio, demonstrating the potential of TTS in related fields.

Neural Source is a waveform model that improves speech generation speed through direct training using spectral training criteria and stochastic gradient descent [6]. The source module, filter module, and condition module work together to produce fast and high-quality speech generation. However, despite the simplified training process, as a non-autoregressive model, it still requires careful parameter tuning to ensure stability. Additionally, the modular design and spectral training criteria demand significant computational resources, particularly GPU support, indicating that the speech generation process still needs further improvement.

In terms of parallel speech generation, FastPitch is an exemplary model. FastPitch introduces pitch forecast for improving the standard of synthesized speech without increasing computational overhead and maintains efficient parallel generation capabilities [7]. However, experimental results have shown that it does not surpass some strong baseline models like WaveNet [7]. To address this issue, FastDiff employs time-aware position-variable convolution stacks and can produce voice with great fidelity without Mel-spectrograms, accomplishing a cutting-edge Mean Opinion Score (MOS) of 4.28 [8]. This model has an extremely fast sampling speed, requiring only four iterations to synthesize high-quality, high-fidelity speech, achieving 58 times real-time speed on a V100 GPU [8]. However, fundamentally, the reduced computational cost is minimal, and due to the characteristics of the diffusion model, its stability still needs improvement. Differences can be found in below Figures about their architectures.

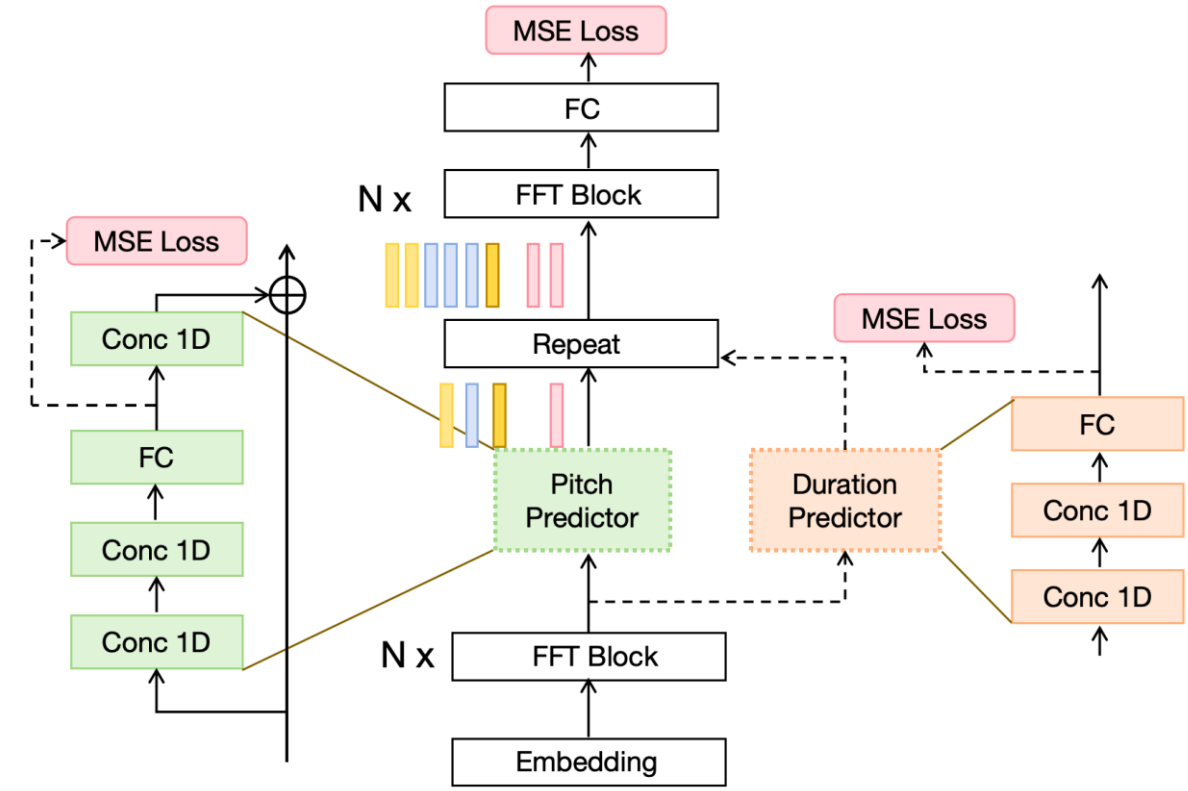

Fig. 3 Architecture of FastPitch [7]

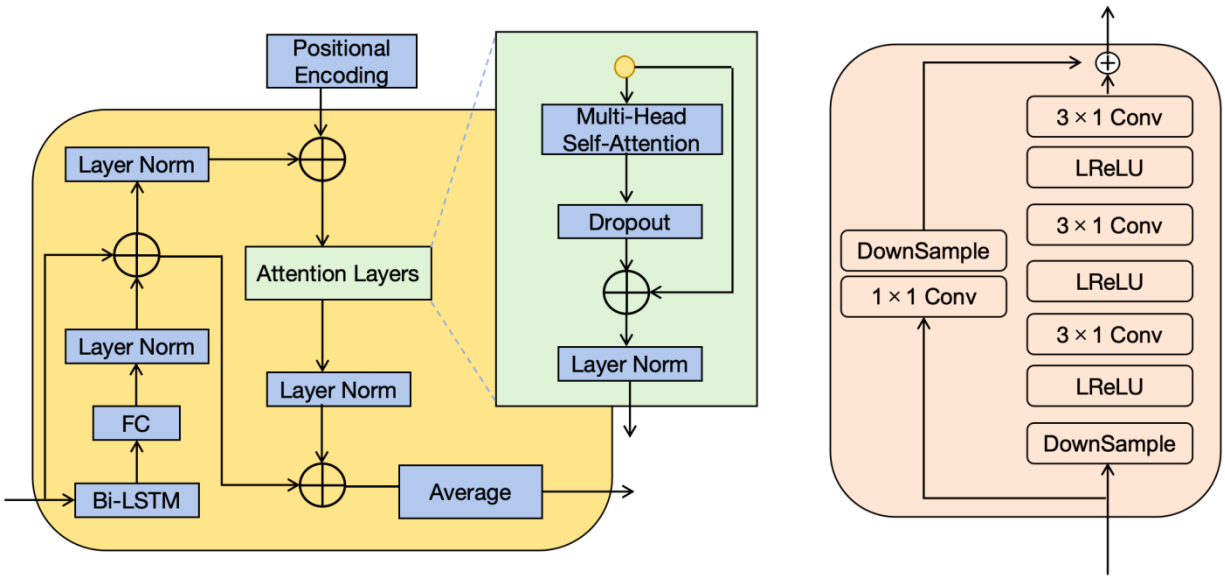

Fig. 4 Architecture of FastDiff [8]

From the above three representative models, it is evident that key challenges include: How to reduce computational cost and complexity? How to improve model stability? And how to ensure high-quality output with minimal data dependency? These are pressing issues that need to be addressed in the current research landscape. It is gratifying that one significant research implication of the aforementioned various speech synthesis techniques is the enhancement of intelligent speech recognition capabilities. On one hand, advanced speech synthesis technology can generate natural multi-speaker synthesized speech, which can replace expensive manually transcribed speech, thereby saving costs. Although synthesized speech cannot completely replace human speech at present, increasing the amount of synthesized speech data can still maintain good recognition performance while reducing the reliance on human speech data [9]. On the other hand, synthesized speech can

generate speech data with diverse pronunciations and styles, effectively improving the robustness of speech recognition systems to noise and variability, thus performing better during evaluation or inference [9].

2.2. Decoupling and Synthesis of Speech

2.2.1. Decoupling of Speech

To enhance the flexibility and control of Text-to-Speech (TTS) systems, recent research has focused on improving pitch adjustment and achieving speaker information disentanglement without significant loss of speech content. CONTENTVEC is one of the effective methods developed in this area. It is based on the HuBERT framework and incorporates innovative modifications, including teacher disentanglement, student disentanglement, and teacher prediction conditioning [11]. Specifically, CONTENTVEC achieves effective speaker disentanglement by removing speaker information from the teacher labels and applying a regularization loss to the students' speech representations [11]. Additionally, it inputs speaker information into the masked prediction task, thereby reducing the need for the speech representation to encode speaker information [11]. These improvements enable TTS systems to better adapt to various application scenarios and user needs while preserving content. However, despite CONTENTVEC's remarkable performance in speaker disentanglement and its ability to achieve this without significant content loss, the removal of speaker information may still lead to slight content degradation. By gradually increasing the weight of contrastive loss during training, CONTENTVEC effectively enhances the disentanglement of speaker information and demonstrates superior performance in maintaining the integrity of content information [11].

2.2.2. Speech Synthesis

In recent years, speech synthesis technology has made notable improvements, especially with the adoption of deep learning methods that provide robust support for generating high-quality, natural-sounding speech.

Speech synthesis models like SpeechX and GAN-TTS have made breakthroughs in multitask learning and high-fidelity speech generation. SpeechX uses a neural encoder-decoder language model combined with multitask learning, enabling speech enhancement and editing in various environments, demonstrating the model's versatility and robustness. GAN-TTS, nevertheless, uses Generative Adversarial Networks to generate high-fidelity speech that is on par with state-of-the-art models and has effective parallelization capabilities [12]. Overall, modern speech synthesis technology continues to optimize the architecture and algorithms of generative models. By introducing new evaluation metrics and diverse datasets, the quality and efficiency of speech synthesis have been significantly improved, paving the way for broad applications in scenarios such as voice interaction, virtual assistants, and language learning. However, both models require substantial computational resources and data, have complex architectures, and their performance can be affected by background noise and other factors [12,13].

Voicebox, representing a new generation of speech generation models, employs non-autoregressive flow matching technology, completely revolutionizing traditional speech synthesis methods. Voicebox is trained on over 50,000 hours of raw speech data, covering audio snippets from various real-world scenarios, ensuring its exceptional performance across different language environments [14]. Unlike traditional autoregressive models that generate speech incrementally, Voicebox can quickly generate complete speech segments. This enables it not only to achieve zero-shot text-to-speech synthesis in both single and cross-language contexts but also to efficiently handle complex tasks such as noise removal, content editing, and style transformation [14]. In practical applications, Voicebox exhibits high flexibility and adaptability. It can automatically infer the style and content of audio without any preset labels, generating contextually appropriate speech [14]. This label-independent feature endows Voicebox with stronger generalization capabilities in handling various speech generation tasks. In contrast to current state-of-the-art models, Voicebox shows significant

improvements in speech intelligibility and audio similarity. Due to its non-autoregressive generation mechanism, Voicebox's inference speed is several times faster, making it particularly outstanding in real-time processing applications [14]. This substantial performance boost is mainly attributed to the latest flow matching method used during training [14]. This method optimizes the neural network parameters, allowing the model to transform from simple probability distributions to complex audio distributions, thereby achieving high-quality speech generation. Moreover, Voicebox can balance between speech quality and computational efficiency by adjusting the number of generation steps. This feature not only provides a strong technical advantage in the field of speech synthesis but also opens new development directions for future speech generation technologies. Notably, the model has developed an efficient classifier to discern between produced and actual speech, significantly preventing the possibility of abuse [14].

In summary, despite the significant achievements in the efficiency, multifunctionality, high-quality output, and flexibility of speech synthesis, the related model architectures remain complex and require substantial computational resources, greatly limiting their application.

3. Multimodal

3.1. Speech-to-Gesture Synthesis

Speech-to-gesture generation has advanced significantly in the last few years, with researchers exploring various innovative methods and technologies. Early studies focused on predicting appropriate gestures corresponding to speech signals. One notable study introduced the use of a denoising autoencoder neural network to learn low-dimensional representations of human motion. This approach, utilizing a motion encoder and decoder, retains the most important elements of human posture variations while removing irrelevant changes, resulting in more concise and representative motion representations [15]. The principal benefit of this technique is its capacity to successfully capture the main features of gestures and align them with speech signals, facilitating natural speech-to-gesture translation. Unlike traditional methods, this technique does not rely on manually labeled data but instead employs a data-driven approach to generate gestures, offering greater flexibility and broader application potential. Additionally, this study delved into the impact of different dimensions of motion representations and speech features on the generation outcomes, finding that medium-dimensional representations and the use of MFCC speech features yielded the best results [15]. However, due to limitations in the training dataset, this approach may lack sufficient diversity and complexity in generated gestures in certain intricate scenarios.

In subsequent research, scientists have further expanded the scope of speech-to-gesture generation by proposing a new approach that leverages multimodal contextual information. This study integrates various input modalities, including text, audio, and speaker identity, to synchronously encode and decode multimodal data, producing gestures that are highly aligned with the speech's topic and cadence [16]. The key innovation in this approach lies in its introduction of adversarial training mechanisms and the proposal of a new quantitative evaluation metric—Fréchet Gesture Distance (FGD)—to measure the distribution differences between generated and real gestures [16]. This method not only generates more natural and diverse gestures but also allows for the creation of gesture styles specific to different speaker identities, significantly enhancing the model's adaptability and application flexibility [16]. However, due to the high complexity of this approach, which involves the integration of multimodal data and adversarial training, it requires substantial computational resources and training data, posing potential challenges in practical applications.

Overall, these studies illustrate the technological evolution in the field of speech-to-gesture generation from single-modality to multimodality approaches. From the initial low-dimensional motion representations to the current methods incorporating multiple modalities, these innovations have not only improved the naturalness and accuracy of gesture generation but also expanded its capabilities in terms of diversity and personalization [15,16]. Despite the differences in implementation details and technical focus, these approaches collectively propel the technology towards a more intelligent

and realistic future, laying a solid foundation for applications in speech interaction and virtual character animation.

3.2. Personalized Speech Generation

In recent years, personalized speech generation has made significant progress, focusing on producing speech that is not only natural and high-quality but also capable of capturing the unique characteristics of different speakers. Meta-StyleSpeech represents a major breakthrough in this field. The model aims to generate high-quality speech while effectively adapting to the style of new speakers using minimal reference audio [17]. In order to match the style vector retrieved from reference speech, the model incorporates Style-Adaptive Layer Normalization (SALN), which modifies the gain and bias of text inputs [17]. The architecture of StyleSpseech is shown as follows.

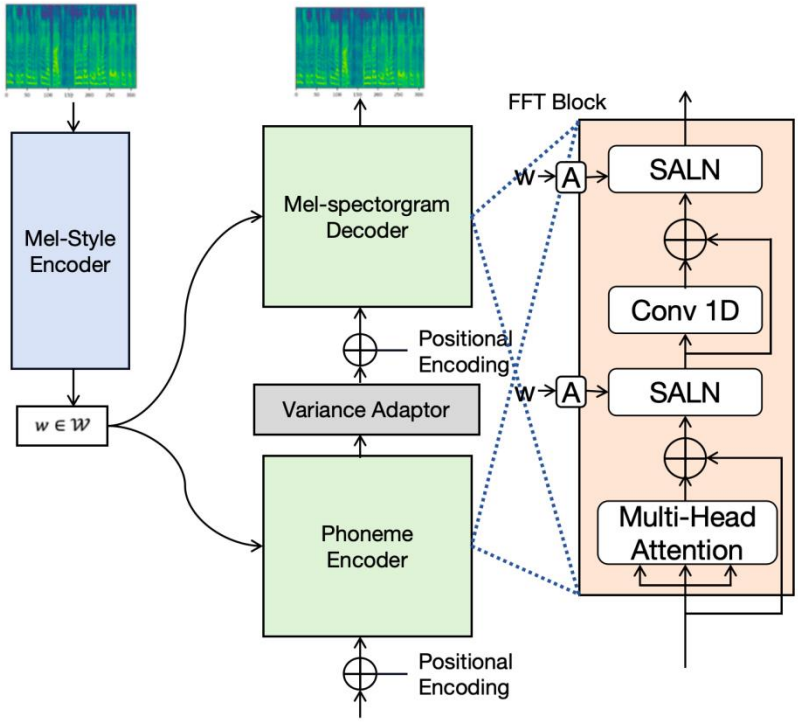

Fig. 5 Architecture of Meta-StyleSpeech [17]

This method allows the model to summarize discourse in the manner of the intended speaker even from a short audio clip, providing great flexibility and adaptability. However, relying on a reference audio sample means that the attributes and traits of the output speech are highly dependent on the quality of the input sample, which can be a limitation if high-quality reference audio is not available. Meta-StyleSpeech extends its capabilities by incorporating two discriminators trained using episodic training and style prototypes [17]. This enhances the model's ability to generalize to unseen speakers. By simulating one-shot adaptation scenarios, the model can quickly adapt to new speakers with minimal data [17]. This capability is particularly useful for applications requiring rapid adaptation to new voices, such as personalized virtual assistants and speech synthesis for individuals with unique vocal characteristics. The primary flaw in this approach is the increased complexity of the model, requiring significant computational resources for training. Handling multiple styles and speaker identities can pose challenges in real-time applications.

Fine-grained emotion strength transfer is another noteworthy method for creating emotional speech synthesis. Without the need for manual labels or reference recordings, our approach creates sophisticated and contextually appropriate emotional expressions in synthetic speech by combining sentence-level emotion categories with phoneme-level emotion strength representations [18].The architecture of it is shown as follows.

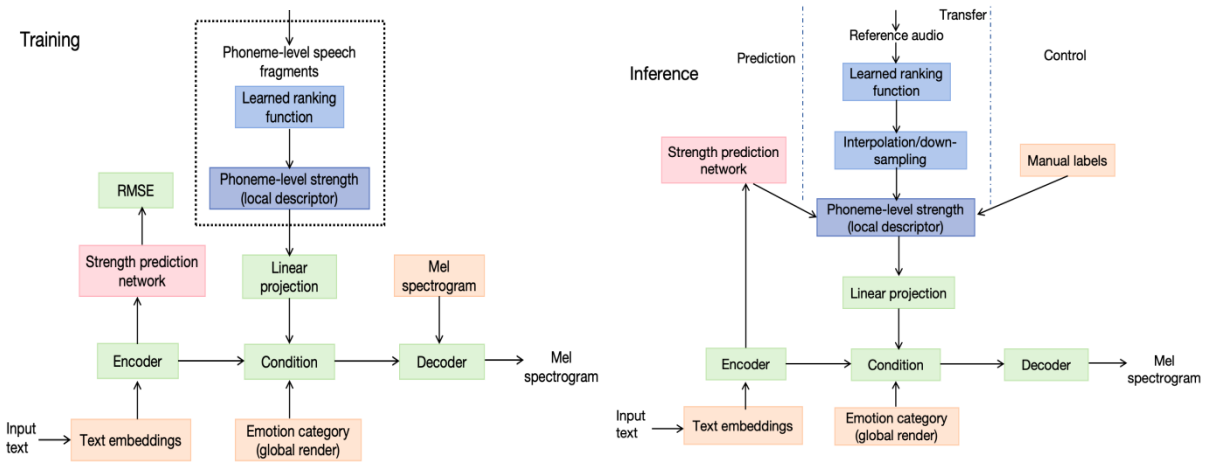

Fig. 6 The system [18]

The ability to control local emotion descriptors makes speech synthesis more personalized and expressive, ideal for applications in entertainment and customer service where nuanced emotional expression is crucial. However, the challenge lies in the need for detailed and accurate emotion strength annotations, which are difficult to obtain and may limit the model's generalizability across different emotional contexts and languages. Another method effectively addresses these limitations by enabling controllable speech synthesis. This approach enhances the learned units using learned speech units as vocoder input, quantized F0 representation, and global speaker embeddings, allowing for the separation of content, prosody, and speaker identity from speech. This method can achieve an encoding rate of 365 bits per second, significantly outperforming baseline methods [18].

A brand-new direct speech-to-speech translation technique called Translatotron 2 suggests that preserves the speaker's voice across different languages without the need for speaker segmentation or additional speaker representations [19]. This method ensures that the synthesized speech retains the original speaker's identity, which is crucial for applications where speaker consistency and privacy are important, such as voice cloning and translation services [19]. The advantage of this approach lies in its simplicity and effectiveness in maintaining speaker identity, even during speaker turns. However, this method requires large multilingual datasets, and managing multiple language models can pose challenges, especially in resource-limited environments.

Overall, personalized speech generation has made significant strides in creating more natural, flexible, and contextually appropriate speech outputs. The advancements in models like Meta-StyleSpeech, fine-grained emotion synthesis, and Translatotron 2 highlight the potential of personalized speech synthesis in various applications [17-19]. While these models offer impressive capabilities, they also face difficulties like the requirement for high-quality reference data, high computational resource requirements, and the complexity of managing multiple styles and speaker identities. As research continues to evolve, overcoming these challenges will be key to further enhancing the effectiveness and accessibility of personalized speech generation technologies.

4. Relevant Datasets

- LibriTTS Dataset: LibriTTS is a large-scale public corpus containing 585 hours of English speech data from the LibriVox project. The dataset features recordings from multiple speakers and includes high-quality audio and annotation information [10].
- LJSpeech Dataset: LJSpeech contains 13,100 English speech clips recorded by a single female speaker, totaling approximately 24 hours of audio [3,10].
- VCTK Dataset:VCTK is a diverse speech dataset featuring multiple speakers, encompassing approximately 44,000 speech clips from 109 speakers, encompassing a variety of accents and speech styles [3,10].

- Conversational Spanish-to-English Dataset: This dataset focuses on Spanish-to-English speech translation and includes rich conversational speech data for evaluating cross-language speech translation models [19].
- CoVoST 2 Dataset: CoVoST 2 is a multilingual speech translation dataset containing translation pairs from French, German, Spanish, and Catalan to English, used for researching cross-language speech translation [19].
- IEMOCAP Dataset: IEMOCAP is an emotional speech dataset comprising 10 sessions recorded by 10 speakers, covering a variety of emotional categories such as anger, happiness, sadness, and more [18].
- SEMAINE Dataset: The SEMAINE dataset contains approximately 5 hours of emotional speech dialogue recordings, featuring a variety of emotional expressions from multiple speakers [18].
- TED Talk Dataset: This dataset contains videos, audio, and text data from TED Talks, featuring a rich variety of speakers and contextual backgrounds, covering a wide range of topics and language styles [16].
- Human3.6M Dataset: Human3.6M is a large-scale human motion capture dataset, documenting a variety of daily activities and human motions [16].
- MNIST Dataset: MNIST is a benchmark dataset containing images of handwritten digits, widely used for evaluating machine learning and deep learning models [15].
- CIFAR-10/100 Datasets: CIFAR-10/100 are small image datasets containing images of 10/100 different object classes, commonly used for research in image classification tasks [15].
- ImageNet Dataset: ImageNet is a substantial picture repositorycontaining millions of images across over 1,000 object categories, used for deep learning training and evaluation [15].
- MVTec AD Dataset: MVTec AD is an industrial defect detection dataset, including images of various defect types for anomaly detection and data augmentation research [15].
- Retinal OCT Dataset: This dataset is used for diagnosing retinal diseases and includes a large number of optical coherence tomography (OCT) images of the retina [15].
- PASCAL VOC Dataset: PASCAL VOC is a standard dataset for object detection and image segmentation, providing rich annotation information [15].
- COCO Dataset: COCO is a large-scale image dataset containing extensive data for object detection, image segmentation, and captioning tasks [15].

5. Conclusion

Speech synthesis and personalization under monomodal and multimodal conditions have become highly significant research directions in recent times. In recent years, research on personalized text-to-speech conversion through waveforms and the generation of associated gesture features has increased significantly and achieved notable progress. This article provides an overview of the latest advancements in this field, offering a detailed comparison of various methods and models, highlighting their advantages and disadvantages. It also addresses the pressing issues in current research and discusses potential future directions that could drive the development of monomodal and multimodal speech synthesis. These discussions not only offer valuable reference and guidance for future research but also provide strong support for the practical application of speech synthesis technologies.

6. Abbreviations

This manuscript employs the following abbreviations:

TTS	Text-to-Text Speech
GAN	Generative Adversarial Network
	Variational Autoencoder
	Short-Time Fourier Transform
	Mean Opinion Score

VAE

STFT

MOS

DL

SPSS

CNN

RNN

LSTM

MFCC

COCO

References

[1] Yamamoto R, Song E, Kim J M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6199-6203.

[2] Tan X, Chen J, Liu H, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[3] Yang G, Yang S, Liu K, et al. Multi-band melgan: Faster waveform generation for high-quality text-to-speech[C]//2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 492-498.

[4] Polyak A, Adi Y, Copet J, et al. Speech resynthesis from discrete disentangled self-supervised representations[J]. arXiv preprint arXiv:2104.00355, 2021.

[5] Bińkowski M, Donahue J, Dieleman S, et al. High fidelity speech synthesis with adversarial networks[J]. arXiv preprint arXiv:1909.11646, 2019.

[6] Wang X, Takaki S, Yamagishi J. Neural source-filter-based waveform model for statistical parametric speech synthesis[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5916-5920.

[7] Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6588-6592.

[8] Huang R, Lam M W Y, Wang J, et al. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis[J]. arXiv preprint arXiv:2204.09934, 2022.

[9] Rosenberg A, Zhang Y, Ramabhadran B, et al. Speech recognition with augmented synthesized speech[C]//2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019: 996-1002.

[10] Ning Y, He S, Wu Z, et al. A review of deep learning based speech synthesis[J]. Applied Sciences, 2019, 9(19): 4050.

[11] Qian K, Zhang Y, Gao H, et al. Contentvec: An improved self-supervised speech representation by disentangling speakers[C]//International Conference on Machine Learning. PMLR, 2022: 18003-18017.

[12] Mustafa A, Pia N, Fuchs G. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6034-6038.

[13] Wang X, Thakker M, Chen Z, et al. Speechx: Neural codec language model as a versatile speech transformer[J]. arXiv preprint arXiv:2308.06873, 2023.

[14] Le M, Vyas A, Shi B, et al. Voicebox: Text-guided multilingual universal speech generation at scale[J]. Advances in neural information processing systems, 2024, 36.

[15] Kucherenko T, Hasegawa D, Henter G E, et al. Analyzing input and output representations for speech-driven gesture generation[C]//Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019: 97-104.

[16] Yoon Y, Cha B, Lee J H, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity[J]. ACM Transactions on Graphics (TOG), 2020, 39(6): 1-16.

[17] Min D, Lee D B, Yang E, et al. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation[C]//International Conference on Machine Learning. PMLR, 2021: 7748-7759.

[18] Lei Y, Yang S, Xie L. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis[C]//2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 423-430.

[19] Jia Y, Ramanovich M T, Remez T, et al. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation[C]//International Conference on Machine Learning. PMLR, 2022: 10120-10134.