Artificial Intelligence Speech Synthesis Based on the Deep Learning

Sihang Li[a]
Beijing-Dublin International College at BJUT, Beijing University of Technology, China

Keywords: Artificial Intelligence, Speech Synthesis, Deep Learning.

Abstract: Speech synthesis is one of the most popular topics in machine learning, and it aims to generate an expressive voice that can satisfy the demands of different fields. This survey introduces the technology of generating speech based on deep learning. Besides, it reviews the development of autoregressive frames (represented by Transformer TTS) as well as non-autoregressive (represented by Fastspeech) in terms of speech synthesis. Autoregressive frame strengths in generating expressive speech while non-autoregressive tend to efficiently generate that. Moreover, extra refined programs based on the above are also included. It contains an Autoregressive Acoustic Model with Mixed Self-attention and Lightweight Convolution(AAMSLC), Autoregressive Diffusion Transformer(ARDiT), RoubuTrans, FastSpeech 3, FastPitch, ProbSparseFS, LinearizedFS, LightTTS and so forth. These techniques represent current cutting-edge advances in the field of speech synthesis. The purpose of this passage is to provide a systematic knowledge review for beginners in the field, which helps them to better understand the latest developments in speech synthesis technology while providing new ideas for future research and applications. .

1 INTRODUCTION

Speech synthesis is a way that machine can generate speech through given text(also called text-to-speech), which is widely deployed to human-computer natural language interaction (HCM), Intelligent customer service, voice navigation, assistance for the blind, audiobooks, and game voice-overs. Speech synthesis is a technique that originated in the 18th century with mechanical synthesis devices. Later, it developed from the Vocoder electronic synthesizer in the 20th century, to the DECtalk resonance peak synthesis technology. Then it went to the unit splicing-based synthesis and statistical parameter-based synthesis techniques. Though speech synthesis has experienced rapid development and has made great progress, the synthesized speech is still not as good as it could be. Besides, high labor costs, Complex parameter adjustment, and problems such as distorted speech and stiff intonation were still the blocks of applying speech synthesis. As technology advances, The deep neural network-based speech synthesis method overcomes the above problems better and generates high-quality speech through a simple and flexible process, which has led to significant development of speech synthesis technology (Tan et al., 2021).

The autoregressive framework is one of the most mature types in end-to-end speech generation systems. It generates sequences on the principle of conditional probability, which means the next sequences will be generated based on what has already been produced. In general, the Autoregressive frame has a sequential, one-way synthesis pattern. The advantage of this framework is that it can capture the sequential dependencies of the sequences well and thus exhibit strong coherence and logic, showing natural content, timbre, rhythm, emotion, and style (Tang et al., 2023). Whereas, its low synthesis speed, slow training speed, and high training costs result in discomfort with faster interaction scenarios such as face-to-face conversations or conversations.

In recent years, the non-autoregressive framework has been proposed as one of the natural language processing frameworks based on the autoregressive framework, which is distinctly different from the latter one. The differences have been shown in rapid, parallel generation of sequences and shorter training

[a] https://orcid.org/0009-0001-9057-6550

Li, S.
Artificial Intelligence Speech Synthesis Based on the Deep Learning.
DOI: 10.5220/0013702500004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 591-595
ISBN: 978-989-758-765-8

time. However, the generation series of non-autoregressive is lower in accuracy and fluency compared with the autoregressive framework due to its inherently relatively simple structure and weak order dependence.

This paper investigates the speech synthesis literature in recent years. It reviews current developments in this field in terms of autoregressive and non-autoregressive frameworks, with Transformer and FastSpeech as the classical frameworks respectively. Besides, all frameworks will be evaluated by performances, qualities, training and calculation costs, and so forth. This paper aims to help practitioners of speech synthesis understand its development, provide fundamental knowledge of speech generation techniques, and realize the current problems in this field.

2 SPEECH SYNTHESIS

Deep learning for speech generation has been a research frontier in the field in recent years. In the process of generation, texts are passed to the TTS front end to extract the feature of pronunciation and rhyme (i.e. the basic language feature). Then it would be received by the acoustic framework and transformed into an acoustic feature (most commonly the Mel-spectrum), which is the most essential part of the text-to-speech process. The acoustic framework can also be divided into the autoregressive and non-autoregressive framework. Lastly, The vocoder converts the Mel-spectrogram into sound waves and eventually synthesizes speech.

2.1 Transformer and its improved versions

As proposed by Vaswani et al., Transformer TTS frameworks are typical of autoregressive frameworks (Vaswani et al., 2017). The principle is that the text is numericized and passed into the Transformer framework, which consists of several encoders and decoders. The numericized text is first via the encoder, which has a significant sub-attention Layer. Its self-attention mechanism allows each word to be encoded taking into account the influence of other words in the sentence as they are encoded. After passing all the encoders, texts will be passed on to decoders where texts re-experience the sub-Attention Layer. The final output of the Mel-spectrogram can be translated to approximate human speech. Noticeably, the texts have been passed to a number of attention layers. This is called a multi-head attention mechanism, which enables the framework to extract different semantic information. The advantages of Transformer TTS are parallel training, fast training, and good processing results. However, it also encountered problems such as slow synthesis, poor fine-grained control, and instability such as possible missing information between word positions over long distances.

For the past few years, there has been constant improvement and innovation in the Transformer TTS framework. As proposed by Zhao, AAMSLC Replacing partial self-attention operations with a small number of convolutional operations showed a powerful performance (Zhao, 2023). The results of his experiment have revealed that the hybrid frameworks of different structures can be trained and reason more efficiently(36 percent in training efficiency and 95 percent rise in that of reasoning). Moreover, it can also avoid a considerable number of redundant operations. However, the Introduction of Convolution has increased the complexity of frameworks as well as the demand for hardware. Lastly, when dealing with multiple other languages, its performance may not reach the same level as Chinese.

Liu et al. proposed ARDiT frameworks as another revised version that relies on the Transformer framework's decoder to encode audio as a sequence of vectors in continuous space, rather than the traditional sequence of discrete symbols (Liu et al., 2024). Therefore generate natural and high-quality speech. In addition, it exhibits strong performance in terms of sample-free speech generation. Whereas the training complexity and computational cost of this method are relatively elevated and consume more computational resources.

The last framework of autoregressive to be introduced is RobuTrans, which has been developed based on the robustness of the Transformer (maintaining stability and output capacity when encountering disruptions or unknown contingencies) (Li et al., 2020). This framework addresses the instability of the existing neural TTS when dealing with anomalous text by converting the input text into sequences containing phonemes and rhythmic features with duration-based hard and pseudo-non-causal attention mechanisms and removing positional embedding to achieve high-quality and stable speech synthesis. Studies have demonstrated that RobuTrans achieved natural and robust superiority over other

frameworks. Besides, no additional teacher frames are needed to achieve good compositing results.

2.2 FastSpeech and its improved versions

A typical example of a non-autoregressive is FastSpeech (Ren et al., 2019). It is a Transformer-based end-to-end TTS system. It has been proposed to tackle problems such as slow reasoning and low controllability. The process of its generation is first phoneme sequences are input. Then they are converted to embedding format. After that, they are processed through multiple encoder-like FFT Blocks. Subsequently, it is passed to a Length Regulator that predicts the length of the Mel-spectrogram and controls the rhythms to improve frame controllability. Finally, the final Mel-spectrogram is output again via multiple FFT blocks. In addition, FastSpeech introduces knowledge distillation to improve the quality of audio. Knowledge distillation simplifies the distribution of output data and enhances the framework's one-to-many language processing problem. In general, FastSpeech 2 significantly improves speech generation speed when compared with Transformer. However, it shows a disadvantage in the quality of generation and has drawbacks such as complex and time-consuming knowledge distillation.

FastSpeech 2 is one of the revised versions proposed by the same team(Ren et al., 2020). It is an advanced non-autoregressive text-to-speech (TTS) model designed to address the limitations of its predecessor, FastSpeech, while enhancing synthesis quality and efficiency. By eliminating the complex teacher-student distillation pipeline, FastSpeech 2 directly trains on ground-truth mel-spectrograms, avoiding information loss and simplifying the training process. To alleviate the one-to-many mapping challenge in TTS, the model introduces variance information such as pitch, energy, and precise phoneme duration extracted through Montreal Forced Aligner (MFA). Notably, pitch prediction is improved via continuous wavelet transform, which models pitch variations in the frequency domain for higher accuracy. Additionally, FastSpeech 2s extends the framework by enabling fully end-to-end text-to-waveform synthesis, bypassing intermediate mel-spectrogram generation and achieving faster inference speeds. Evaluations on the LJSpeech dataset demonstrate that FastSpeech 2 surpasses autoregressive models like Tacotron 2 in voice quality (MOS: 3.83 vs. 3.70) and reduces training time by threefold compared to FastSpeech. The model ' s ability to integrate variance control (e.g., adjustable pitch and energy) enhances prosody customization while maintaining naturalness. By combining simplified training, enhanced variance modeling, and end-to-end capabilities, FastSpeech 2 advances the practicality of high-quality, real-time speech synthesis with robust controllability.

FastSpeech provides an excellent basic framework for others. FastSpeech-based FastPitch is a parallel text-to-speech model built upon FastSpeech, designed to enhance speech expressiveness and quality by explicitly predicting fundamental frequency (F0) contours (Łańcucki, 2021).. The architecture utilizes two feed-forward Transformer stacks: one processes input tokens, while the other generates mel-spectrogram frames. By conditioning on F0 values predicted at the granularity of input symbols, the model resolves pronunciation ambiguities and improves speech naturalness. During inference, users can intuitively adjust predicted pitch values to modify prosody, enabling natural voice modulation while preserving speaker identity. FastPitch achieves exceptional synthesis speed, with a real-time factor exceeding 900 times for mel-spectrogram generation on GPUs, outperforming autoregressive models like Tacotron 2. Evaluations on the LJSpeech-1.1 dataset demonstrate superior Mean Opinion Scores (4.080) compared to Tacotron 2 (3.946) and multi-speaker models such as Flowtron. Unlike FastSpeech 2, which predicts frame-level F0, FastPitch operates at the symbol level, simplifying interactive pitch editing without compromising quality. The model also supports multi-speaker synthesis through speaker embeddings, delivering state-of-the-art performance. By combining high-speed parallel synthesis with flexible prosodic control, FastPitch advances applications in expressive and real-time speech generation, offering practical advantages in both quality and usability.

ProbSparseFS and LinearizedFS are also two newly proposed FastSpeech-based TTS frameworks that significantly improve inference speed and memory efficiency while maintaining speech quality through efficient self-attention mechanisms and compact feed-forward networks(Xiao et al., 2022).

The last framework of non-autoregressive to be introduced is LightTTS, which is a lightweight, multi-speaker, multi-language text-to-speech (TTS) system(Li et al., 2021). It achieved fast speech synthesis of different languages or codes by deep

learning. Notably, Compared to traditional attention-based autoregressive frameworks, LightTTS employs non-autoregressive generation, which considerably improves the synthesis speed and drastically reduces the framework parameters. Experiments showed that LightTTS can generate Mel-spectrogram 2.5 times faster than FastSpeech. However, the number of participants decreased by a factor of 12.83. It is comparable to real speech in terms of naturalness and similarity, demonstrating its promise for a wide range of applications in multilingual environments.

3 LIMITATIONS AND DIRECTIONS FOR DEVELOPMENT

Research in the field of Speech Synthesis has made remarkable progress in recent years, leading to significant advancements in the naturalness and expressiveness of generated speech. These developments have far-reaching implications across various industries, including entertainment, education, healthcare, and human-computer interaction. However, despite these achievements, several challenges remain unresolved, indicating that the field still has a long way to go before achieving truly human-like and versatile speech synthesis systems.

One of the most pressing issues is the latent problem of privacy leakage, which arises from the misuse of large datasets required to train sophisticated frameworks. The collection and utilization of voice data often occur without explicit consent or proper anonymization, raising ethical and legal concerns. Additionally, the quality of voice data in these datasets is highly inconsistent, leading to uneven training effects for specific frameworks. While screening datasets to select appropriate resources could mitigate this issue, the process is often costly and time-consuming, posing a significant barrier to efficient model development.

Another critical limitation is the current framework's inability to achieve effective one-to-many mapping. Existing systems still rely heavily on additional inputs, such as intonation, accent, rhythm, and emotional cues, to generate high-quality speech. This dependency not only increases the complexity of the frameworks but also demands substantial human effort in data annotation and preprocessing. Furthermore, the application of explicit modeling techniques, while effective in certain scenarios, significantly raises the cost and time required for both training and deploying these frameworks. On the other hand, implicit modeling, though more efficient, often falls short in terms of performance and versatility.

The generation speed of current frameworks, despite noticeable improvements, remains a bottleneck for real-time voice interactions. While recent advancements have accelerated the synthesis process, the computational demands of high-quality speech generation still hinder seamless real-time applications, particularly in scenarios requiring low latency and high responsiveness.

Looking ahead, this paper proposes several directions for future research to address these challenges. First, the disclosure and misuse of personal privacy can be mitigated through stricter legislation and regulations governing the collection and use of datasets. Establishing clear guidelines and ethical standards for data usage will be crucial in building trust and ensuring compliance. Second, conducting rigorous pre-screening of datasets to ensure quality and consistency will enhance their utility and improve training outcomes. Third, optimizing algorithms and exploring innovative ways to combine different frameworks or integrate frameworks with convolutional techniques could pave the way for simultaneous improvements in generation speed and speech quality. Finally, leveraging advanced networks such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) could enhance the one-to-many mapping capabilities of speech synthesis systems, enabling more flexible and expressive voice generation.

In conclusion, while significant strides have been made in speech synthesis, addressing the remaining challenges will require a multidisciplinary approach that combines technological innovation, ethical considerations, and regulatory frameworks. By focusing on these areas, the field can move closer to achieving truly human-like, efficient, and privacy-conscious speech synthesis systems.

4 CONCLUSIONS

This paper discussed deep learning-based speech synthesis techniques focusing on autoregressive and non-autoregressive frameworks. Autoregressive frameworks like Transformer TTS are capable of

generating more natural and expressive speech but slowly and costly. Non-autoregressive frameworks such as FastSpeech increase generation speed, it is lack accuracy and fluency. This paper also introduces various improved versions that have been explored and refined to increase synthesis speed, reduce frame parameters, improve speech quality, and adapt to multilingual environments. Despite significant progress, it still faces problems such as dataset abuse, high framework complexity, one-to-many mapping, high framework training cost, and slow generation speed. Future research can explore different frameworks or the combination of frameworks and convolution to improve the generation speed and quality while referring to encoders, global style labeling, VAE, GAN, and other techniques to improve the one-to-many mapping capability of frameworks. Ultimately, autoregressive and non-autoregressive research continues to promote speech synthesis technology in the direction of more efficient and natural, providing a technical basis for intelligent interaction and multi-scene applications.

REFERENCES

Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023). A Survey of Expressive Speech Synthesis, Big Data Research, 9.6: 53-71.

Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan N., G., Lukasz, K., & Illia, P. (2017) Attention is All You Need, Advances in neural information processing systems, 30: 5998-6008.

Zhao Wei. (2023). Research on deep neural network acoustic modeling for efficient speech synthesis (PhD dissertation, Zhejiang University).https://link.cnki.net/doi/10.27461/d.cnki.gzjdx.2023.000815doi:10.27461/d.cnki.gzjdx.2023.000815.

Liu, Z., Wang, S., Inoue, S., Bai, Q., & Li, H. (2024). Autoregressive Diffusion Transformer for Text-to-Speech Synthesis. arXiv preprint arXiv:2406.05551.

Li, N., Liu, Y., Wu, Y., Liu, S., Zhao, S., & Liu, M. (2020, April). Robutrans: A robust transformer-based text-to-speech model. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 8228-8235).

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). Fastspeech: Fast, robust and controllable text to speech. Advances in neural information processing systems, 32.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

Łańcucki, A. (2021, June). Fastpitch: Parallel text-to-speech with pitch prediction. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6588-6592). IEEE.

Xiao, Y., Wang, X., He, L., & Soong, F. K. (2022, May). Improving fastspeech tts with efficient self-attention and compact feed-forward network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7472-7476). IEEE.

Li, S., Ouyang, B., Li, L., & Hong, Q. (2021, June). Light-tts: Lightweight multi-speaker multi-lingual text-to-speech. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8383-8387). IEEE.