Analysis of Speech Synthesis Technology: From Deep Learning to Airflow Modeling

Tian Xie[a]
Chongqing Nankai Secondary School, Chongqing, 400069, China

Keywords: Deep learning; Speech Synthesis Technology; Airflow Modeling.

Abstract: With the advancement of artificial intelligence, speech synthesis technology has been widely applied across multiple fields. Deep learning-based speech synthesis has gained significant attention due to its ability to automatically learn complex acoustic features, greatly improving speech fluency and naturalness. This paper reviews deep learning-based speech synthesis technology, with a particular focus on its applications in falsetto and voice transformation tasks. By exploring the principles of human speech production and the development of speech synthesis technology, this paper analyzes the advantages and limitations of current deep learning models and proposes an innovative method that integrates articulatory organ parameters with acoustic parameters. Furthermore, the paper discusses the potential of airflow simulation in physical modeling, especially its application prospects in generating personalized voices and handling voice transformation and falsetto tasks. Finally, this paper outlines future research directions, including optimizing deep learning models, integrating physical modeling techniques, and fostering interdisciplinary research, aiming to advance speech synthesis technology towards greater personalization and richer emotional expression..

1 INTRODUCTION

Speech synthesis technology refers to the process of converting input text sequences into speech output with high naturalness, high audio quality, and rich expressiveness through appropriate prosodic processing and specific synthesizers. This enables computers and related systems to produce natural and fluent speech similar to human voices(Zhang, 2014).

In recent years, with the continuous advancement of artificial intelligence (AI) technology, speech synthesis has become a crucial field in modern computer science. Its applications span various industries, from intelligent assistants and automatic speech recognition to virtual character dubbing in the entertainment industry. In these applications, generating natural and fluent speech is essential for enhancing user experience and system performance. As technology has progressed, speech synthesis has evolved from articulatory synthesis, which relies on mechanical physical modeling of speech production, to formant synthesis, which is based on source-filter models and resonance peak weighting. Due to technological limitations, both of these early methods could only generate relatively simple speech sounds. Later, with rapid advancements in computer hardware, waveform concatenation emerged and gradually matured as a more effective method.

Despite significant advancements in speech synthesis technology, challenges remain in generating highly natural and personalized speech. These challenges primarily involve synthesis speed, timbral naturalness, and speech diversity. To overcome the limitations of traditional methods, researchers have introduced Hidden Markov Models (HMMs) for parametric statistical speech synthesis and leveraged deep learning algorithms to address difficulties in discovering and modeling acoustic features. While deep learning models have achieved breakthroughs in the fidelity and coherence of speech generation, they still face notable limitations in specific tasks such as falsetto (fake voice) synthesis and voice conversion, particularly in timbral transformation, emotional expression, and voice imitation.

This paper aims to review deep learning-based speech synthesis techniques, with a particular focus on their applications in falsetto and voice conversion

[a] https://orcid.org/0009-0003-6503-5036

Xie, T.
Analysis of Speech Synthesis Technology: From Deep Learning to Airflow Modeling.
DOI: 10.5220/0013702300004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 581-585
ISBN: 978-989-758-765-8

tasks. Specifically, we will explore how deep learning models simulate human timbral variations, especially in cases involving emotional fluctuations and voice mimicry. Additionally, we will analyze the limitations of current techniques in these tasks and discuss potential directions for future improvements.

This paper is organized as follows: The second part introduces the principles of human vocalization and the development of speech synthesis technology; the third part discusses the advantages of unconditional speech synthesis and analyzes the limitations of unconditional speech generation technology in dealing with tasks such as voice changing and false voice; the fourth part proposes ideas based on computer physical modeling and airflow simulation and summarizes the advantages and disadvantages of this method; the fifth part summarizes the whole paper and puts forward prospects.

2 DEVELOPMENT OF SPEECH SYNTHESIS TECHNOLOGY

2.1 Principles of Human Phonation

Human phonation is the result of the coordinated operation of multiple speech organs, which can be broadly categorized into respiratory organs, phonatory organs, articulatory organs, and resonating cavities. The primary steps of phonation involve the contraction of the lungs generating airflow, which then passes through the trachea into the larynx. The vocal cords located in the larynx vibrate according to the speed and pressure of the airflow. The airflow is further modulated by passing through the oral and nasal cavities, which act as resonators shaped by the structure and openings of these cavities. The frequency of vocal cord vibrations determines the fundamental pitch of the sound, while adjustments in the tension, shape, and vibration speed of the vocal cords can regulate pitch and volume.

2.2 Mechanically and Electronically Based Speech Synthesis

Research on speech synthesis dates back to the late 18th century, with the earliest synthesis methods relying on physical devices to simulate human phonation. These methods attempted to replicate speech by mimicking the movement of speech organs and modeling airflow to produce simple sounds. With advancements in electronics, early physical synthesis devices were gradually replaced by the source-filter model. This method conceptualizes the speech generation process as a source simulating glottal states, which excites a time-varying digital filter that characterizes the resonant properties of the vocal tract. It primarily uses waveform superposition to simulate the vocal cords, oral cavity, and other organs (Jing, 2012). Voiced sounds are generated using a pulse generator, while unvoiced sounds are produced by a noise generator, and after passing through a vocal tract filter and lip radiation process, the final speech signal is synthesized (Zhang, 2016). These early mechanical and electronic methods could only generate simple speech, with limited accuracy in vocal tract simulation, making them impractical for real-world applications.

2.3 Phoneme Concept and Waveform Concatenation

Although human languages are diverse, all are composed of phonemes. Phonemes are the fundamental units of synthesized speech, with all words and sentences formed by concatenating multiple phonemes. Speech synthesis can be achieved by assembling pre-recorded speech units from a speech database to generate complete utterances. This concatenation method retains the original speaker's timbre to the greatest extent, providing high naturalness and intelligibility, reaching an acceptable level for human listeners.

Despite the high-quality synthesis achieved through concatenation, the generated speech often sounds artificial and rigid, with issues in prosody at concatenation points. The most direct solution to these problems is to record a large-scale speech corpus in various contexts, addressing the discontinuities at unit boundaries found in traditional waveform concatenation methods. However, such a large corpus requires significant storage space and is time-consuming to produce. Additionally, an efficient algorithm is needed to select the correct speech units for concatenation from the database.

This approach involves two main processing modules: text processing and acoustic processing. The front-end module, responsible for text processing, converts input text into a symbolic phonetic description–determining what sounds to produce and how to produce them. The back-end module, responsible for acoustic processing, transforms these symbolic descriptions into the acoustic features of speech signals (Zhu, 2009).

2.4 Hidden Markov Model and Parametric Synthesis

Although waveform concatenation and traditional synthesis methods partially address prosodic discontinuities, further optimization is needed. In practice, traditional concatenation methods often suffer from unstable synthesis quality and low robustness due to incorrect unit selection. To solve this issue, a superior speech unit selection algorithm is required to accurately predict the acoustic parameters corresponding to a given text under different conditions.

The Hidden Markov Model (HMM) is a double-stochastic process where the specific state sequence is unobservable, but its transition probabilities are known. That is, the state transitions are hidden, while the observable events are random functions of these hidden transitions (Jing, 2012).

From a statistical perspective, human speech production is also a double-stochastic process. The brain, based on expressive needs, organizes language according to grammatical rules and generates a series of unobservable commands to control speech organs, ultimately producing observable acoustic parameters. This speech generation process is similar to the description of HMM, making HMM a cornerstone method in statistical parametric speech synthesis(Wu, 2006).

The HMM-based model primarily involves two phases: training and synthesis. The training phase consists of five steps: model initialization, HMM training, context-dependent HMM training, decision tree-based training, and duration modeling. Once the model is trained, it can generate state sequence feature vectors from input text. These vectors are then processed by a filter to convert them into speech signals (Wu, 2006). The HMM-based modeling approach offers greater flexibility, does not require an extensive speech corpus, and significantly reduces the time needed for model construction compared to traditional methods. As a result, it is more suitable for lightweight and embedded platforms.

3 SPEECH SYNTHESIS TECHNOLOGY BASED ON DEEP LEARNING

3.1 Speech synthesis technology based on deep learning

Early speech synthesis technology represented by HMM-based speech synthesis will inevitably destroy the fine structure of natural speech spectrum while using statistical parameters. Moreover, due to the limitation of computing power, it can only consider the influence of one or two adjacent phonemes, resulting in the discard of potential meaningful information in the previous text, causing information loss(Pan, 2021).

In end-to-end speech generation, the system consists of two parts: acoustic model and a vocoder. The acoustic model realizes the temporal alignment of text and speech, and the vocoder restores the output of the acoustic model into a speech waveform(Zhang, 2021). The essence of speech generation is to simulate sound through a series of acoustic parameters. Acoustic parameters are a kind of complex data and are not easy to model manually. Deep learning can learn more useful features by building a machine learning model with many hidden layers and big data training, which just solves the problem that acoustic parameters are not easy to model and select features manually. DNN is a common model for modeling acoustic parameters. In training, the minimum mean square error criterion is usually used to train the DNN model, and the model parameters are continuously adjusted to minimize the error between the predicted acoustic parameters and the target acoustic parameters. In synthesis, after extracting text features, the trained DNN model is used to predict the acoustic parameters and the duration information provided by other systems (such as the PSOLA algorithm), and then input into the vocoder to obtain synthesized speech(Zhang, 2020).

3.2 Processing and limitations of voice change

Under ideal conditions, the above-mentioned deep learning-based speech synthesis algorithms can generate fluent and natural speech more accurately. However, in real life, acoustic parameters do not only include prosodic parameters. The same person can not only make one voice. The volume, pitch, voice emotion, and switching between true and false voices will affect the acoustic parameters. Moreover, emotions will also affect people's control over various parts of the body, including the vocal organs. These special control rules under emotional conditions will produce special emotional speech parameter changes. In the synthesis system, when the range of variation of the rhythmic acoustic parameters is large, the synergy between the articulatory organs will have a greater impact on the speech, and the spectrum and filter can no longer be simply treated as completely independent parameters (Wang, 2013). The sole use of rhythmic acoustic parameters will lead to a single and unnatural result in the actual application of speech synthesis. In this regard, new parameters need to be introduced to solve such problems.

3.3 Integration of Articulatory and Acoustic Parameters

In HMM-based speech synthesis, the primary modeling targets are acoustic parameters and observable speech features. However, acoustic parameters are not the only way to represent speech characteristics; articulatory parameters also serve as effective descriptors. As mentioned earlier, human speech production involves multiple vocal organs, and various technologies such as X-ray imaging and ultrasound can capture articulatory organ parameters (Wang, 2008). Compared to acoustic parameters, articulatory parameters provide more fundamental speech representations, exhibit smoother and more gradual changes, and demonstrate better robustness, making them more suitable for HMM-based modeling.

4. PHYSICAL MODELING, AIRFLOW SIMULATION, AND SPEECH SYNTHESIS

4.1 Reflections and Innovations in Physical Methods

Scientists have long recognized that human speech production results from the coordinated operation of multiple vocal organs. However, early physical devices were relatively simplistic and ineffective due to the lack of advanced computing technology. With the development of computer hardware and the maturation of computational physical simulation, it has become theoretically feasible to use computer modeling to simulate airflow within vocal organs. Furthermore, by adjusting parameters according to individual physiological differences, it is possible to generate more precise and personalized speech synthesis.

4.2 Advantages and Limitations

This method has the following advantages: most human vocal organs are similar, so simulating different human voices only requires fine-tuning of parameters, and does not require completely independent modeling for all situations; when simulating voice change and false voice, rapid construction can be performed based on the experience of physiological observations, and there is no need to completely rebuild the voice change or false voice library; the vocal organs are relatively stable and are rarely affected by acoustic noise and environmental noise. They can identify erroneous data to a certain extent and are more robust; physical modeling can well simulate the actions of humans when switching phonemes during pronunciation, and when synthesizing speech, the connection between different phonemes is more natural;

At the same time, this method also has many problems. The following are some of the most representative ones: human pronunciation is not only determined by physiological structure but acquired learning and pronunciation habits are likely to affect the results. Therefore, when using physical models, some acoustic parameters need to be introduced; if the pronunciation organs of each audio provider are scanned, it will cost a lot of costs and resources, and it is necessary to develop an algorithm that can infer the structure of human pronunciation organs based on audio; although there is no need to model the original sound, this method will cost a lot of computing power in the physical simulation of airflow.

4 CONCLUSIONS

This paper reviewed deep learning-based speech synthesis technologies, with a particular focus on advancements in falsetto and voice transformation tasks. With the rapid progress of deep learning, traditional speech synthesis methods have gradually been replaced by more complex and precise deep neural networks (DNNs). By leveraging large datasets and deep learning models, speech synthesis systems can generate high-quality, natural, and fluent speech, significantly improving the user experience. However, despite these breakthroughs, existing technologies still face considerable challenges in handling tasks such as timbre variation, emotional expression, and voice imitation.

Firstly, current deep learning models lack sufficient flexibility in voice transformation tasks, especially in timbre adjustment and emotional variation. Although strategies that integrate articulatory and acoustic parameters have been introduced, further research is needed to address the complexity of acoustic feature modeling. Secondly, physical modeling and airflow simulation methods have shown promising potential, particularly in accurately capturing the natural transitions and variations of vocal organs, thus improving speech realism and robustness. However, physical modeling still faces challenges such as high computational resource consumption and difficulties in accurately capturing individual differences.

To overcome these challenges, future research may focus on the following directions: On the one hand, enhancing the adaptability of deep learning models in voice transformation and falsetto tasks by incorporating finer-grained emotional and timbre parameters to achieve more precise personalized speech synthesis. On the other hand, optimizing physical modeling techniques to explore how to achieve realistic and personalized voice simulations with lower computational costs. Additionally, interdisciplinary research − such as integrating neuroscience, acoustic engineering, and artificial intelligence − may bring breakthroughs in speech synthesis technology.

In conclusion, while speech synthesis technology has made significant strides across various fields, achieving high-quality voice transformation and falsetto synthesis still requires substantial technological innovation and interdisciplinary collaboration. As computing power continues to improve and algorithms become more refined, future speech synthesis systems are expected to make remarkable advancements in personalization, emotional expressiveness, and speech diversity, further driving the development of intelligent voice interaction technologies.

REFERENCES

Jing, X., Luo, F., Wang, Y., 2012. A review of Chinese speech synthesis technology. In Computer Science, 39(S3), 386-390.

Pan, X., Lu, T., Du, Y., et al., 2021. A review of speech synthesis and conversion technology based on deep learning. In Computer Science, 48(08), 200-208.

Wang, R., Dai, L., Hu, Y., et al., 2008. A new generation of speech synthesis technology based on acoustic statistical modeling. In Journal of University of Science and Technology of China, (07), 725-734.

Wang, Y., 2013. Research on key technologies in speech synthesis system. Beijing, Tsinghua University.

Wu, Y., 2006. Research on speech synthesis technology based on hidden Markov model. Hefei, University of Science and Technology of China.

Zhang, B., Quan, C., Ren, F., 2016. A review of speech synthesis methods and development. In Small and Micro Computer Systems, 37(01), 186-192. DOI:10.20009/j.cnki.21-1106/tp.2016.01.035.

Zhang, X., Xie, J., Luo, J., et al., 2021. A review of deep learning speech synthesis technology. In Computer Engineering and Applications, 57(09), 50-59.

Zhang, Y., 2020. Research on speech synthesis algorithm based on deep neural network. Xi'an, Xidian University. DOI:10.27389/d.cnki.gxadu.2020.002643.

Zhang, Z., 2014. Research on Chinese speech synthesis based on deep neural network. Beijing, Beijing Institute of Technology.

Zhu, W., 2009. Linguistic computational models in speech synthesis: current status and prospects. In Contemporary Linguistics, 11(02), 159-166+190.