Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

A Review of Applications of Speech Synthesis Technology

Wenhao Xiong *

Department of Information Science and Engineering, East China University of Science and Technology, Shanghai, China

* Corresponding Author Email: 23013139@mail.ecust.edu.cn

Abstract. Currently, the development of visual technologies and applications is much more advanced than that of speech, but as both speech and vision are equally important and attractive, their potentials should be comparable. In order to develop speech technology and prove its application potential, this paper presents current status of speech synthesis technology in four applications: spoken language education, digital music, virtual character, and language protection and dissemination, and then points out its potential in different stages of spoken language learning, music generation similar to current picture generation, virtual characters that move people in ways other than through singing, and new ways of voice protection and dissemination both directly and indirectly, and finally discusses the balance of the development of speech synthesis technology. By analyzing several potential applications of speech synthesis technology from application areas, this paper not only proves the development potential of speech, but also shows the way for subsequent research to find innovative inspiration from application areas.

Keywords: Speech synthesis technology; Application; Prospect.

1. Introduction

Speech is the most commonly used way for humans to communicate with the outside world [1], and it is the maximum way for humans to obtain external information second only to vision, which is of high research value. Compared with the more complex visual technology, which is still in the booming development period, speech technology is relatively more mature. However, this does not mean that speech technology has reached a bottleneck. On the contrary, there are a large number of potential applications waiting for it. What is missing is eyes who can spot these potential 'gaps'.

Speech technology contains speech synthesis and speech recognition. Speech synthesis technology refers to converting text information to speech information, which allows people to obtain information conveyed by computer through hearing. The fundamental theories of modern digital signal processing have laid a solid foundation for speech synthesis technology. In recent years, with the rapid development of deep learning, speech synthesis technology has made great progress [2]. Shortly after the emergence of sequence-to-sequence networks incorporating the attention mechanism, end-to-end speech synthesis method has become a research hotspot again, which has raised the speech synthesis technology to an upper level. This method is able to directly convert text into speech waveforms without intermediate phoneme or syllable conversion, which has higher naturalness and flexibility. It also allows acoustic and language models to be utilized for better technical results [3].

In the field of furniture and transportation, intelligent voice assistants such as Apple Siri, Amazon Alexa use deep-learning-based speech synthesis technologies, such as Tacotron 2 and WaveNet model. Through end-to-end architecture and efficient vocoder, they can respond naturally and smoothly to user instructions [4]. In the field of broadcasting, with the help of speech synthesis technology, virtual hosts can be applied to news broadcasting, customer service, pavilion information introduction, etc., which not only has smooth and emotional voice close to real broadcast hosts, but also has the advantages of low cost, high broadcasting efficiency, and high broadcasting accuracy. At the same time, it also provides a new idea for the development of traditional broadcasting hosts [5]. In the field of medicine, systems based on speech synthesis technology can help special users or patients such as illiterate people, language learners, the elderly, and people with learning disabilities

or vision loss by reading text, and are also changing traditional medical treatment and promoting transformation of medical system [6].

Although speech synthesis technology has been used in a wide range of fields, such as education, entertainment, healthcare, transportation, etc. and is steadily growing in each field, there is still a lot of untapped potential.

The following examples introduce the gaps in speech synthesis technology from the application aspect. For the field of intelligent interaction, in complex contexts, the problem that speech synthesis is prone to stiff and incoherent needs to be solved. In real-time interaction, the delay of speech synthesis caused by large model volume and insufficient arithmetic power needs to be solved. In education and security, speech synthesis technology involves collecting and processing of a large amount of speech data, so data privacy and security become especially important.

The following examples introduce the gaps in speech synthesis technology from the technical aspect. Currently, most of the mainstream and commonly used speech synthesis models have the problem of large volume and datasets, resulting in speech synthesis not being able to be performed in time for certain application scenarios or run smoothly on small and medium-size computing devices. As a result, many tasks, such as zero-shot speech synthesis [7], multi-language speech synthesis [8], and applications, such as personalized content creation [9], are limited.

Scholars are aware of the strong potential of speech synthesis technology, so they are not only optimizing speech synthesis technology in traditional application areas, such as education, security, and healthcare, while developing various applications of speech synthesis technology in emerging or special areas, such as virtual reality, accessibility, culture, etc., but are also exploring more novel relevant technologies and applications, such as visual speech that can be used to drive virtual characters, perform face forgery detection, etc [10].

For further development of speech synthesis technology, this paper explores various present situations of speech synthesis technology from its applications, in order to better explore its role and performance in applicaations, to discover its improvement, and to propose its new development directions.

This paper chooses to focus on four application areas: spoken language education, digital music, virtual character, and speech protection and dissemination. The paper provides an overview of current status of speech synthesis technology in these four applications and then analyzes the needs and potentials to give suggestions on the direction of the development of speech synthesis technology.

2. Present Situations of Application of Speech Synthesis Technology

2.1. Spoken Language Education

The application of speech synthesis technology in the field of spoken language education has made remarkable development and become an important tool to enhance teaching efficiency and learning experience.

Speech synthesis technology can provide learners with pronunciation demonstrations in different languages, different genders, different accents, different emotions, etc., meeting diverse learning requirements. Combined with speech synthesis technology, human-computer interactive teaching enhances the interest and participation of learning by practicing dialogues with students through virtual characters.

Table 1 below gives some specific applications of speech synthesis technologies in spoken language education.

Table 1. Specific applications of speech synthesis technology in speaking instruction.

Application	Examples of specific applications	Functional embodiment of speech synthesis technology
Translation	Youdao	Speech synthesis technology provides pronunciation for learning when translating sentences, paragraphs, etc..
Word memorization	Duolingo, Babbel, Rosetta Stone	Speech synthesis technology plays standard pronunciation of words for guidance.
Speaking test simulation	TalkMe	Speech synthesis technology provides pronunciation for virtual teachers to conduct simulated test dialogue exercises.
Conversation practice	Hi Echo, Love Password	Speech synthesis technology provides users with pronunciation demonstrations, enabling users to practice conversations with virtual teachers anytime, anywhere, covering a wide range of life and academic scenarios.

2.2. Digital Music

Speech synthesis technology is becoming more and more widely used in the field of digital music. The essence of this kind of music lies in the deconstruction, in-depth analysis and learning of massive music data through precise algorithms, so as to construct style models with relatively certain aesthetics. Its characteristics compared to real people are as follows.

Music and songs can be generated easily and quickly. Users can select specific styles to generate music and songs, or synthesize various timbres of vocals, chords, sound effects, etc., such as chorus, harmony, etc. For professionals, speech synthesis technology can be used to create music demos for quicker and more convenient auditioning. Speech synthesis technology can also be used for music mixing and post-editing, by processing and modifying the audio to make it richer and more diverse [11].

For example, MusicLM model released by Google can convert text signals into audio clips. It can create music according to different places, times or requirements, and its samples are composed of long melody of specified music genres, music atmosphere and even specific instruments. It can also infer additional melody sections based on the sound of the user humming, whistling, or playing an instrument, adjust the type of instrument and playing strength, and even create a continuous music improvisation.

2.3. Virtual Character

The application of speech synthesis technology has given rise to the formation of many virtual characters, ranging from beloved characters in games, physical robots with specific service functions to virtual characters in interactive, query and guidance systems. Virtual idols are definitely in the most named ranks among them. They have attracted fans worldwide through their impressive singing voice, unique image, interactive ability, and other charms. The most famous among them are shown in Table 2 below.

Table 2. Application of speech synthesis techniques to virtual idols.

Virtual idol name	Applications of speech synthesis technology
Hatsune Miku	Through VOCALOID technology, Japanese virtual idol Hatsune Miku creates songs in a variety of genres, from pop to electronica, classical to rock. Hatsune Miku has also released audio libraries in multiple languages, including Japanese, English, and Chinese, making her highly influential globally.
Lottie	Through VOCALOID technology, Chinese virtual idol Lottie creates Chinese songs full of Chinese cultural elements. Lottie incorporates more multimodal technology in her concerts and events, such as fusion with motion capture data, which allows the virtual character's voice to be naturally synchronized with her movements.

2.4. Language Protection and Dissemination

Speech synthesis technology plays an important application value in the field of language protection and dissemination, especially in the digital preservation and inheritance and cross-cultural dissemination of endangered languages.

In terms of language protection, speech synthesis technology records and reproduces the phonetic features of endangered languages through digital means, which provides strong support for long-term protection of languages. In addition, speech synthesis technology is also able to model the phonetic features of endangered languages, so that even if the language is lost, it can be reproduced by technical means [12]. In terms of language dissemination, whether in specialized endangered speech dissemination, cultural inheritance, or in cross-language learning platforms, small language synthesis projects, speech synthesis technology plays direct and indirect effects of language and culture dissemination.

Table 3 below shows some application cases of language protection and dissemination to reflect the role of speech synthesis technology.

Table 3. The role of speech synthesis technology in language protection and dissemination projects.

Project name	The role of speech synthesis technology
China Language Resources Protection Project	The project uses speech synthesis technology to leave "sound specimens" for endangered languages, and combines the multimodal corpus platform to digitize and preserve the speech, text and cultural resources of endangered languages [13].
Nanjing Digital Landmark Project	The project provides audio guide services for local languages through speech synthesis and AR technology, which not only preserves local language and culture, but also promotes its application in modern cultural and tourism scenarios.
IMS Toucan	The project is an open-source multilingual synthesis platform capable of synthesizing speech in over 7000 languages. Its extensive language support provides a technological foundation for the protection and dissemination of endangered languages [14].

3. Prospects for the Application of Speech Synthesis Technology

3.1. Spoken Language Education

Language learners are generally less competent in speaking than in listening, reading and writing for the following reasons. The conditions and resources for speaking training are several times higher than those for the other three, and the time spent by the majority of language learners on speaking training is much lower than that for the other three. At the same time, the best effect of oral training is to communicate with native speakers in real-life scenarios, which is both the ultimate goal of oral learning and the reason why it is so difficult to train well in speaking, as the other three languages are better able to simulate real-life situations for training. However, speaking is the most urgent need for training in real scenarios among these four. This shows that speech synthesis technology has great potential in teaching spoken language.

Speech synthesis technology can provide assistance to language learners at all stages of learning, and specific future directions can be developed as shown in Table 4 below.

Table 4. What speech synthesis technology can do for language learners at different stages of learning.

Learning stage	Stage description	Applications of speech synthesis technology
Beginning	Pronunciation learning	For the speech synthesis pronunciation of words, phrases, sentences, paragraphs, etc. in translation software and vocabulary memorization software, it adds more contextualized emotion and voice intonation, as well as personalized pronunciation adjustment functions, so that learners can learn more realistic and standard pronunciation.
Skilled	Pronunciation practice	In order to better train speaking, communicating with real people in a real environment is the best choice, but only a very small number of learners can have such a condition, so virtual dialog is especially important. In addition to the vigorous development of virtual teachers such as TalkMe, it is also possible to carry out multimodal fusion, so that the virtual teacher can see the surrounding things, play some images or add more virtual characters to increase the authenticity and complexity of the scene, and enhance the learning effect.
Advanced	Environmental integration	Proficiency in speaking does not mean that you can integrate into the native environment of the language you are learning. Commonly used colloquial expressions, popular verbal phrases, and speech intonation of characteristic scenarios are difficult to access in conventional learning, which poses a higher challenge for speech synthesis technology.

3.2. Digital Music

Music plays an extremely important role in people's lives, emotions, social interactions, and personal development. A vast majority of people are very fond of, dependent on, or even fanatical about music. Therefore, any application of speech synthesis technology in the field of digital music has a great possibility to be extremely attractive to a large group of people. Compared to image generation, which is hot right now, music generation can be said to be quite cold. However, music may have the same potential as image generation or even more. Therefore, by analogy, the current development of image generation can be used to reason about the high potential development directions of music generation. Table 5 below shows some examples.

Table 5. Image generation reasoning about music generation development directions.

Image generation		Music generation		Specific descriptions of music generation development directions
Test to Image	Prompt	Test to Image	Prompt	Can guide musical synthesis by cue words that can make demands on vocal content, melody, harmony, instrumentation, and emotional expression.
	Model		Model	Can choose different style models as the base, such as rock, pop, hip-hop, electronic, heavy metal, folk, etc.
	Parameter		Parameter	Can be stylization, variety, wonder, seamless music, etc. similar to Vince's chart.
Image to Image	Local adjustion	Image to Image	Local adjustion	Can make certain modifications to aspects of the music by cueing in localized parts of the music while keeping the overall flow of the music.
	Image expansion		Music expansion/ contraction	Can partially expand or contract music while ensuring the unity of the overall style of music.
	Image beautification		Music beautification	Can beautify the human voice and instrumental sound in the music, such as reducing the broken voice, changing the tone, and fixing some small mistakes. If it can do real-time beautification, its application scene is bigger.

3.3. Virtual Character

Nowadays, virtual idols are the most popular kind of virtual characters among most population, which shows their potential. However, among them, not many are famous now. Therefore, the reasons for

this is worth exploring. Based on the results of the analysis, the speech synthesis technique should be improved.

From the fame of virtual idols such as Hatsune Miku and Lottie, it can be seen that virtual characters can move people's hearts through their singing voices and thus receive such a wide range of popularity and love, so why can't they do the same in other aspects? Table 6 below shows the development directions of virtual characters in aspects other than singing, but undoubtedly they are all closely related to speech synthesis technology. In the future, through further analysis or exploration of their prospects, speech synthesis technology also needs to be upgraded accordingly.

Table 6. Potential of VMs in all directions.

Potential direction	Introduction
Transmit	Shape specific virtual announcers, who can also have entities. They have their own set of acting styles like real people, and can serve as newscasters, radio announcers, feature piece commentators, etc., or even more demanding positions such as tournament commentators, interviewers, and other positions that require both specialized knowledge and resilience.
Helper	Shape specific virtual assistants, who can also have entities. These assistants can be housekeepers, communicating with people of different identities, such as house owners and guests, executing commands for the use of electrical appliances, and answering simple questions such as searching for items, asking for information, and so on. These assistants can also be companys' "brain", answering different questions and delivering information for people in different positions, facilitating centralized management. Even cell phones, laptops, or access to websites and transportation systems can have virtual assistants. The MOSS in Wandering Earth, Jarvis in Iron Man, and the Red Queen in Resident Evil are the target forms of virtual assistants mentioned here.
Emotional companionship	Shape specific virtual emotional companions, who can also have entities. They can accompany users and communicate with them as if they were family members or friends, and can provide users with emotional value, or relieve the loneliness of specific groups of people, or accompany children as toys or educational tools. At present, Groove X's LOVOT, Hasbro's My Real Baby, etc. all belong to one type of emotional companion, and various types will surely proliferate in the future. This application field is still in its infancy and has unlimited potential.

3.4. Language Protection and Dissemination

There are various ways of language protection and dissemination, some of which are direct and some of which are indirect. Some of these ways are illustrated in Table 7 below.

Table 7. Examples of direct and indirect modes of language protection and dissemination.

	Language protection	Language dissemination
Direct	Establishment of linguistic databases, speech databases, and multimodal corpora.	Publicize in various ways to increase the attention of the community. Establish learning platforms, courses, etc. Establish model areas for the protection of endangered languages.
Indirect		Foreign language education, learning other languages. International exchanges, experiencing different languages and cultural backgrounds. Dialect broadcasting, increasing listeners' exposure to niche languages.

However, for direct methods of language dissemination, they face difficulties such as low social participation, a weak sense of dialect and national language identity, and the speed of preservation failing to keep up with the speed of language extinction. The poor results of traditional propaganda methods, such as television, radio, lectures, and protection demonstration zones, are telling us that we should follow the development of the times and try new dissemination methods that meet the

characteristics of the times. Nowadays, new media, which are interactive, timely and appeal to a wide range of age groups, are the right choice. There are many ways to use new media for language dissemination, especially Internet new media, such as WeChat public number and video number in social media platforms, and CapCut short video and Red notes in short video platforms. And this variety of language dissemination methods, there are voice synthesis technology shadow and use. For example, it can reduce costs and improve efficiency through virtual hosting technology or AI dubbing in editing software. In addition, it is worth noting that recently on Shake, China's tourist attractions, local officials, and even police and military accounts, in some promotional videos, they will edit popular music, movie character lines, etc., and combine them together to form a novel and interesting "music", which makes the previous boring promotional videos become incredibly attractive. This makes the previously boring promotional video very attractive. Similar videos compared to traditional promotional videos have several times or even dozens of times more than the number of likes. If speech synthesis technology can help creators inspire more by generating something never seen before, the effect on language protection and dissemination needs no further explanation. And by adding some appealing dialects to the mix, it can attract more attention.

At the same time, for indirect dialect dissemination methods, if dialects and other minority languages or endangered languages can be added to specific programs on local TV stations, language broadcasts on buses, language control of smart furnitures and other occasions, they can be exposed to more people's vision, which can also play a role in communication and protection. In these occasions, there are more or less potential applications of language synthesis technology waiting to be developed.

4. Discussion

4.1. Development Balance of Underlying Technology and Upper Applications

Speech synthesis technology should find a balance between the development of the underlying technology and the upper level applications. Although this paper demonstrates the path of technological advancement in search of innovation from applications, a single-minded consideration of applications will always be constrained by the existing technological framework, making it difficult to break free. As in "Three Bodies", human beings are constrained by Sophon, which makes science stagnant, but technology is able to develop rapidly until a very high point. However, it can't go further. The same is true for the relationship between the underlying technology and the upper tier applications. If we try our best to develop upper-layer applications, but the underlying technology is not sufficiently supportive, the applications will be limited and it will be difficult for them to develop further; and vice versa, if the underlying technology is too far ahead of its time and the upper-layer applications cannot keep up with it, the technology will have difficulty in realizing its commercial value and social benefits.

4.2. Cross-domain Universality and Difference

Speech synthesis has a number of applications in most domains. However, the specific needs and application scenarios of different domains impose differentiated requirements on speech synthesis technology. This cross-domain versatility and differentiation indicates that the development of speech synthesis technology needs to be customized and optimized for the characteristics of different domains while maintaining the core advantages of the technology. This not only requires technology developers to deeply understand the specific needs of each domain, but also requires the introduction of multidisciplinary cross-cooperation in the process of technology research and development to ensure that speech synthesis technology can better serve the practical applications in different domains.

4.3. Coordination of Privacy Protection and Technology Application

Speech synthesis technology involves the collection and processing of private and even large amounts of sensitive information in health, education, culture and other fields, so its data privacy and security

are critical. The future development needs to find a coordinated path between technology application and privacy protection. On the one hand, technical means such as end-to-end encryption and anonymous identification can be used to ensure the security of data in the process of collection, storage and use; on the other hand, it is necessary to establish strict data management and privacy protection laws and regulations to standardize the scope of application of speech synthesis technology and the way of data use. Only under the premise of ensuring privacy protection, speech synthesis technology can realize wider application in more application fields.

5. Conclusion

This paper introduces the current status of the use of speech synthesis technology in four application areas: spoken language education, digital music, virtual characters, and voice protection and dissemination, and looks forward to the future development direction of speech synthesis technology by analyzing the application's needs for the target audience and the application potential of speech synthesis technology. Speech synthesis technology can be of significant help to learners of spoken language at all stages. Speech synthesis for generating music may flourish like current AI painting. Speech synthesis technology can open up the same applications for virtual characters in the field of broadcasting, assistants, and emotional accompaniment as it has for singing. Speech synthesis technology can directly and indirectly make more innovative contributions to language protection and dissemination. Finally, this paper discusses the balance between the development of underlying technology and upper application of speech synthesis, cross-domain universality and difference, and the coordination of privacy protection and technology application.

This paper analyzes the current situation of the application of speech synthesis technology, from which new development directions can be proposed. However, these applications are only the tip of the iceberg. There is still unlimited development space for speech synthesis technology. Therefore, the focus of this paper is not only to point out some new development directions of speech synthesis technology in the four application areas of spoken language education, digital music, virtual characters, voice protection and dissemination, but also to show that, from the aspect of applications, by carefully analyzing the demand gap and the new potential of the new or original applications, so as to discover the new development direction, which gives the ability to find innovative inspiration and promote the technological progress of the way. Sometimes, an attention of a small detail of life may become an opportunity for the development of a field that seems almost unrelated.

References

[1] Wei W H. Overview and Research Status of Speech Synthesis Technology. Software, 2020, 41(12): 214-217.

[2] Liu Y. Research and implementation of speech synthesis system based on deep learning. Beijing: Beijing Jiaotong University, 2022.

[3] Liu Y F. A review of the application of artificial intelligence in speech synthesis. Big Data and Artificial Intelligence, 2024, 5(1).

[4] Chen C Y. Speech Synthesis Technology: Status and Challenges. ITM Web of Conferences. EDP Sciences, 2025, 73: 02006.

[5] Gao S. The Influence of Speech Synthesis Technology on Traditional Broadcast Host and Its Development Path. Television Technology, 2024, 48(6): 109-111.

[6] Latif S, Qadir J, Qayyum A, Usama M, Younis S. Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. IEEE Reviews in Biomedical Engineering, 2021, 14: 342-356

[7] Fujita K, Ashihara T, Delcroix M, et al. Lightweight Zero-shot Text-to-Speech with Mixture of Adapters. arXiv, 2024.

[8] Wu Z J, Liu D, Li M. Lightweight Language Model for Speech Synthesis: Attempts and Analysis. 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing, 2024: 501-505

[9] Gong C, Wang X, Erica C, et al. ZMM-TTS: Zero-shot Multilingual and Multispeaker Speech Synthesis Conditioned on Self-supervised Discrete Speech Representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[10] Liu L, Sui J P, Ding D, et al. Reasearch progress and prospects of deep learning for visual speech generation. Journal of National University of Defense Technology, 2024, 46(2): 123-138.

[11] Liu Y X. Research on the practical application of AI speech synthesis technology in the creation of multi-person audio drama. Anhui: Anhui University, 2024.

[12] Lin Y Q, Zhang X X. From Endangerment to Empowerment—An Exploration of Multimodal AI Technologies in Linguistic Diversity Conservation and International Communication Strategies Innovation. Modern Linguistics, 2024, 12(06): 520-529.

[13] Zhang Y. The Construction Practice of the Project for Protecting Language Resources of China: in the Case of the Integration of Anhui Dialects into the Teaching of"Modern Chinese". Journal of Wuhu Vocational Institute of Technology, 2023, 25(2): 64-67.

[14] Lux F, Meyer S, Behringer L, et al. Meta learning text-to-speech synthesis in over 7000 languages. arXiv preprint, 2024.