Proceedings of ICADSS 2024 Workshop: International Forum on Intelligent Communication and Media Transformation
DOI: 10.54254/2753-7064/36/2024BJ0048

Human-machine Voice and Human-machine Competition: Practical Landscape, Pattern Changes and Industrial Challenges of Generative AI-Enabled Audiobook Publishing

Yuanxin Wang[1,a], Tingting Wang[2,b,*], Jingyi Fu[1,c], Xinyue You[1,d]

[1]Department of Publishing, Beijing Institute of Graphic Communication, Xinghua Street, Daxing District, Beijing, China
[2]Department of Economics and Management, Beijing Institute of Graphic Communication, Xinghua Street, Daxing District, Beijing, China
a. wangyuanxin007@163.com, b. wang_tt926@163.com, c.3043568186@qq.com, d. 2559852426@qq.com
**corresponding author*

Abstract: At present, generative artificial intelligence has a profound impact on the audiobook publishing industry chain, and exerts unprecedented transformative power in all aspects of audiobook publishing. This paper reviews the specific technical forms of generative artificial intelligence used in the process of audiobook publishing, including text-to-speech, speech recognition, intelligent speech interaction, voiceprint recognition, voice anonymization, etc., and shows the practical picture of how these technologies influence audiobook publishing in different stages of content creation, sound presentation, recording and editing, quality control, terminal playback and readers' feedback. It analyzes the pattern changes of the audiobook publishing industry brought about by generative artificial intelligence, with "human-machine voice" as the core, and then discusses the challenges and reflections faced by the audiobook publishing industry under the trend of "human-machine competition and cooperation" from the aspects of underlying logic, industry pain points, fundamental principle, reading environment and technical ethics. Therefore, generative artificial intelligence technology is like a double-edged sword, audiobook publishers should fully understand its nature, conform to the trend of technology, and use it to enable new quality productive forces of audiobook publishing industry.

Keywords: generative artificial intelligence, audiobooks, human-computer interaction.

1. Introduction

Generative AI refers to a new technology that generates text, pictures, sounds, videos, codes and other content, which share a striking similarity to that created by humans [1], based on algorithms, models and rules. Audiobook publishing industry is one of the fields to accept, integrate and apply Generative AI, which can soon become a powerful weapon to drive it to make a achievement and promote new quality productivity.

DOI: 10.54254/2753-7064/36/2024BJ0048

2. The Technical Forms of Generative AI Applied in Audiobook Publishing

2.1. Text-to-Speech

Briefly, text-to-speech is a technology that input a text and finally output a speech. It is generally divided into two stages: text analysis and speech synthesis. Traditional text-to-speech needs to manually record a large number of speech samples and output a relatively simple speech through complex processing. But with the advent of "AI+TTS", the synthetic speech has become more natural and accurate. It is now widely used in the stage of production in audiobook publishing.

2.2. Speech Recognition

Speech recognition is a technology that converts human speech into machine-readable text or commands. Its core is to convert speech signals into text, which mainly includes several steps such as speech signal preprocessing, speech feature extraction, acoustic and language model training. "AI+SR" can transform audiobooks from sound-led publications to both audio and text publications, which can bring about a more pleasant reading experience.

2.3. Intelligent Speech Interaction

Functions like "hear, speak, understand" are endowed with audio products in a variety of scenarios, mainly used in the stage of playback to reply readers' feedback. Users can change the interaction mode and complete the story plot in dialogue with machine by voice commands [2].

2.4. Voiceprint Recognition

Voiceprint Recognition is a technology that extracts the speaker's voice characteristics and automatically verifies their identity. When the voiceprint matches, it is verified or retrieved successfully, which can not only determine the identity of the speaker, but also confirm whether any audio contents are the same or not. It is mainly applied in the stage of voice presentation.

2.5. Voice Anonymization

Voice anonymization refers to an asynchronous voice anonymization method that changes speech features to prevent machine recognition but retain human perception, that is, the speaker in the original voice is replaced with a fake one, so as to protect the privacy of the speaker [3]. The technique has already been applied to the voice presentation of audiobooks.

3. Practical Landscape of Generative AI in Audiobook Publishing

3.1. Stage of Content Production

3.1.1. Text Production

Generative AI helps human in the following ways. **One is direct creation.** Such as "*Caiyun small dream*", only by giving the beginning of one story within 1000 words can it continue to write the following content. **The second is creation assistance.** It can provide creators with updated information, creative suggestions and writing framework, saving amounts of time used to spend on searching information. **Third, interactive creation.** It can adapt to readers' needs by changing the story setting, adjusting the characters’ sequence, selecting the appropriate structure by interacting with readers. **Fourth is cooperative creation.** It can choose and combine communication elements such as text, sound, picture and video to present works with "audio and visual integration".

DOI: 10.54254/2753-7064/36/2024BJ0048

3.1.2. Sound Production

First is to reproduce and "revive" sound. It is a speech synthesis technology for both the living and the deceased, supporting the creation of readers' own exclusive sound library. For example, Himalayan APP reproduced the voices of human anchors, such as *Childe Xidao* and *Yidao Susu*, to create AI anchors "Xiaodao Xi" and "Xiaodao Su", with 99,000 and 24,000 fans respectively. They have released 50,030 episodes, 8,338 hours and 17,581 episodes, 2,930 hours of work respectively since their launch in May 2022. Their average working time is much shorter than that of the human anchors, while the number of their works has doubled. It also collected the sound samples of the late famous Pingshu performing artist Tianfang Shan into a sound database to make his voice reappears in the world. The total amount of plays has exceeded 100 million times.

The second is to customize sound. Generative AI supports multi-language and dialect personalized sound library customization, which allows the same audiobook to generate different language versions based on different needs. Meanwhile, it can also use multi-style and multi-emotion TTS to combine and match all the language style, emotion, tone, timbre, pitch, speech speed and other elements. It can also intelligently match readers' current listening scene, automatically generate suitable works.

The third is conceal sound. Generative AI can de-identify the speaker, which perfectly fit both the communication ethics and privacy protection ethics in the era of big data. **The fifth is to combine sound.** Generative AI can distinguish dialogues from narrations, extract dialogue roles in multi-character audiobooks, and create a suitable "combination" with different characters.

3.2. Stage of Recording and Editing

3.2.1. AI Recording and Editing

The essence of AI recording is to complete the process of text to speech by using speech synthesis technology. Himalayan News TTS can convert about 3,000 words per minute, which is an efficiency that real people can't imagine. Also, generative AI drives the content editing productivity with technical power. For example, "*Sound clip*" launched 89 kinds of AI timbre, providing creators with rich creative choices in the process of editing.

3.2.2. Character Recognition

Role-playing dialogues often exist in audiobooks. Generative AI can quickly complete the intelligent chapter dismantling and character recognition, and input the text into the multi-emotional support system.

3.2.3. Automatic Sound Effects

Generative AI can extract various sound effects from the sound library and put them at specified locations according to readers' personal preference. Himalaya, for example, uses the speech signal processing algorithm to create intelligent sound effects, so that users can match different sounds for different content and enjoy a better listening experience.

3.3. Stage of Quality Control

AI content review automatically complete the tedious and repetitive parts to assist the reviewers to speed up the audio content review. Reviewers only need to fill in the feedback opinions to obtain a complete hearing report, without complicated manual summary. The average coverage of automatic content filtering in Himalaya increased from 25.7% in 2022 to 72.2% by the end of 2023.

DOI: 10.54254/2753-7064/36/2024BJ0048

3.4. Stage of Broadcast and Readers Feedback

Readers can wake up the device to start reading, get the story plots or learn others' comments by simple voice commands, and also can determine the direction of the story or choose their favorite voice to perform. Generative AI can automatically capture and deeply learn from the "digital traces" left by readers and incorporate them into its own big data models, making the characteristics of reader-centered more stand out than before.

4. Human-machine Voice: Pattern Changes Generative AI Made

4.1. Improvements of Publishing Efficiency

4.1.1. Reduced Creation Threshold

As for text production, it is also possible for AI to create complete new content without human's motivate imagination, creativity or inspiration. **As for sound production,** on one hand, it is no longer confined to the professional performer for a long time recording and broadcasting; on the other hand, the emergence of AI anchors makes the audiobook more diversified.

4.1.2. Speedup of Production Efficiency

Firstly, the production cost is lowered. Generative AI greatly reduced the labor cost of audiobooks which lower the price and the premium. **Secondly, the production cycle is shortened.** The entire traditional production process was about three times longer than the finished product, and the additional time cost further raised the price of audiobooks. Generative AI can complete the production in a few hours at the fastest, much faster than before. **Thirdly, one-stop production is taking shape.** "AI+TTS" forms a closed loop of the whole process from content creation to sound presentation, providing readers with more convenient audio reading services.

4.2. Changes of Content Ecosystem

4.2.1. Change of Content Production Mode

Gradually, AIGC has become a creation mode alongside UGC, PGC and PUGC on major audiobook platforms, and the completion rate and popularity of AIGC works are no less than that of human works. The total duration of AIGC content on the Himalaya APP has exceeded 240 million minutes, equivalent to 401 million hours, accounting for 6.6% of the platform's audio content by the end of 2023.

4.2.2. Change of Content Production Roles

With the advent of generative AI, the content production have transformed from single-player unicast presentation to AI multicast interpretation, from only human creation to human & AI creation.

4.2.3. Better Audio-visual Experience

generative AI bring about more enriched audio resources. Besides, it also generate visual materials such as pictures, text, animation and video, making audiobooks more three-dimensional. It can also freely switch between listening and watching according to the reader's reading scenario.

DOI: 10.54254/2753-7064/36/2024BJ0048

4.3. Reader Experience Optimization

4.3.1. Customized Services

First, audio customization is more refined. Focusing on readers' needs for more refined timbre, diversified voice lines, emotional fluctuations,etc., the synthesized speech can meet the needs of readers. **Second, the scene adaptation is more accurate.** Generative AI can adapts audiobooks to scenes where the Internet of Things exist [4], and judge the scene and match the hidden needs of readers based on their positioning information or behavioral state, so as to automatically provide sound content suitable for different scenes. **Third, the recommended benchmark is more diversified.** Little consideration was given to readers' needs in timbre, rhythm and other aspects of the sound [5]. The platform subsequently produce a more precise recommendation mechanism related to both text and sound.

4.3.2. Immersive Participation

The role of readers is no longer limited to the target audience, but the "creator", "communicator" or "publisher" who can be deeply engaged in the audiobook production, recording and uploading their own voices to achieve sound creation of specific content.

4.3.3. Matching Service

According to statistics, 88.7% of reader inquiries on Himalaya APP in 2023 will be solved by AI-driven intelligent customer service. Such integrated services enable readers to continuously optimize their reading experience.

5. Human-machine Competition & Cooperation: Challenges Caused by Generative AI

5.1. Underlying Logic: Human-machine Relationship

The role of machine in the audiobook publishing process has changed from the device to the communicator, which not only challenges the inherent logic of communication, but also brings out some new publishing concept of audiobooks.

5.2. Industry Pain Point: Copyright Crisis

There is an opaque "black box" operation in the process of data crawling and processing. Therefor it is easy to become a "hotbed" for breeding infringement, and generative AI may further aggravate the copyright problem and cause industry anxiety.

5.3. Fundamental Principle: Content Quality

Firstly, the matching of generated contents to original contents need to be improved. **Secondly,** due to the lowered creation threshold, the content quality is uneven. **Thirdly**, narrative emotional effect of the synthetic voice is not satisfactory.

5.4. Reading Environment: Information Cocoon

Most of the major audiobook platform recommend fantasy, romance and other "popular" theme of online literature for new users, while books with classic content and high cultural value are rarely on the list, which confines readers to be trapped in the information cocoon. This profit-oriented algorithm logic will further worsened the audio reading ecosystem.

DOI: 10.54254/2753-7064/36/2024BJ0048

5.5. Technical Ethics: Identity and Career Crisis

One is the identity crisis. The human-machine boundary is gradually blurred. The other is the career crisis. It is predicted that about 300 million jobs worldwide will be replaced by generative AI, which will unavoidably lead to career anxiety in all walks of life.

6. Conclusion

As a representative new technology, generative AI has highly penetrated into every stage of audiobook publishing by various technical forms, reshaping the process from content production stage to readers' feedback stage, and revolutionizing the audio publishing industry in publishing efficiency, content ecosystem and reader experience. Meanwhile, it brings about industrial challenges in aspects of human-machine relationship, copyright crisis, content quality, information cocoons and technical ethics. Therefore, generative AI is like a "double-edged sword". Only by fully understanding its nature, conforming to its trend should audiobook publishers use it to enable new quality productive forces.

References

[1] *Wang, J. (2023). Artificial intelligence-generated content and its application in book publishing. Communication and Copyright, (10), 48-51. https://doi.org/10.16852/j.cnki/g2.2023.10.001*

[2] *Shen, Y., & Jin, S. (2022). Mechanism and optimization path of audio publishing in the era of intelligent media. Chinese Editors, (11), 86-91.*

[3] *Wang, R., Chen, L., Lee, K. A., & Ling, Z.-H. (2024). Asynchronous voice anonymization using adversarial perturbation on speaker embedding. In Proceedings of Interspeech.*

[4] *Deng, X. (2020). Evolution logic and advanced path of "ear economy". People's Forum, (05), 95-97.*

[5] *Liu, Y., & Gao, Y. (2019). Application of artificial intelligence speech in audiobooks. Published Studies, (11), 35-39. https://doi.org/10.19393/j.cnki/g2.2019.11.008*