Philosophy and Social Science Vol. 1 No. 5, 2024

AIGC Generative Speech Technology: An Examination of Its Communication Paradigms and Evolutionary Reflections

Wenlu Wei[1], Zhichao Song[2,*]
[1]*School of Communication, Weifang University, Weifang, Shandong, China*
[2]*School of Economics and Management, Communication University of China, Beijing, China*
**Corresponding Author.*

Abstract: The emergence of Artificial Intelligence Generated Content (AIGC) represents a new opportunity for the development of intelligent communication. The development of AIGC technology has given rise to new media transformations, creating new content production and dissemination methods. As one of the core areas of the AIGC application, generative speech technology, due to its low cost and simplicity, has been widely used in scenarios such as audiovisual narration, artificial intelligence anchor broadcasting, and audiobook content production. However, the use of AIGC-generated content in practice, due to default program settings and low barriers to entry, leads to the proliferation of homogenization. On the regulatory side, the lack of effective oversight has led to the misuse of technology and the emergence of communication risks and ethical issues in artificial intelligence. Lastly, the decentralized and fragmented communication model has led to the uncontrollable emergence and development of public opinion events, diluting the discourse power of authoritative media, with rumor clarification becoming the main post-event means to clarify public opinion whirlpools. Based on this, this paper starts from the application scenarios of generative speech technology, reflects on the new issues in the field of "artificial intelligence + speech generation" applications, and aims to provide a deeper technological reflection and new development ideas.

Keywords: AIGC; Generative Speech Technology; Communication Paradigm; Media Ecology

1. Introduction

In 1950, Alan Turing proposed the "Turing Test" to determine whether a machine possesses "intelligence," that is, whether the machine can imitate human thinking to "generate" content and then interact with humans [1]. It can be seen that since then, artificial intelligence has been expected to be used for content creation. In 1954, IBM publicly demonstrated a machine translation system at its New York headquarters. In 1957, the first computer-generated music, Illiac Suite, was born. The success of these early concept validations has raised high expectations for the future of artificial intelligence. AI's ability to create new content based on an understanding of data has only emerged after 2020. Since October 2022, with the global sweep of ChatGPT, AIGC has been widely applied in media, scientific research, entertainment, education, and other fields. AIGC, or Artificial Intelligence Generated Content, uses the computer's supercomputing power and intelligent simulation models to generate various images, texts, audio, and video according to user needs. It has quickly become popular in multiple industries and regions around the world. As an important field of AIGC application, generative speech synthesis technology has many applications. The development and stimulation of AIGC have also made significant progress in fields such as voice recognition and voice dialogue, and have been deeply applied in short videos and live broadcasts on the Internet and have had a profound impact on the content production and dissemination of short videos on social media platforms. The development of media technology has also brought new development opportunities and transformation challenges to the media ecology.

2. Overview of Generative Speech Synthesis Technology

Speech synthesis technology (Text to Speech) converts text into natural speech, involving multiple disciplines such as computer science, linguistics, and acoustics. It plays an important role in human-computer communication. Early speech synthesis methods mainly included parametric and concatenative methods [2]. Nowadays, the development of deep learning has greatly promoted the development of generative speech synthesis systems. Large language models such as Audio LDM, WaveNet, and iFlytek Spark, which are used for audio generation, can now quickly convert text content into spoken language, significantly improving the efficiency of audio production. The main applications involve intelligent virtual anchors, short video dubbing, virtual singer performances, and other fields. Generative speech synthesis technology brings human-computer interaction to a new starting point. It will lead to more trust in machines and algorithms and more expectations for communication with machines. Human-computer interaction is becoming increasingly normalized. Intelligent communication is gradually spreading as a mass behavior and will even promote the mediatization of people's daily lives like social media applications in the past.

Generative speech synthesis technology mainly serves the tasks of human-computer interaction. Depending on whether the tasks in the application are specified, dialogue systems can be divided into two categories: Task-oriented Dialogue systems (TOD) and Open-Domain Dialogue systems (ODD). Specifically, task-oriented dialogue systems focus on completing tasks and solving specific problems [3]. Open-domain dialogue systems are usually data-driven and aim to chat with humans without specific tasks or domain restrictions, such as ChatGPT. Depending on the different application scenarios of generative speech synthesis technology, task-oriented dialogue systems are mainly applied to video and audio dubbing narration scenarios, virtual host voice broadcasting scenarios, audiobook content production, web pages, live broadcast room reception guidance, and other scenarios. Open-domain dialogue systems are mainly used in sports events; game live broadcasting intelligent commentary, artificial intelligence emotional robots, and other scenes [4]. Both types of dialogue systems rely on the learning and generation of the corpus, but the degree of reliance is slightly different.

Nowadays, with more and more "popular" AI production tools represented by Chat-GPT entering the lives of ordinary users, their low cost and simple operation have enabled users to generate a large amount of creative content, which will break the previous content creation process and greatly enhance content production productivity [5].

3. Application Scenarios of Generative Speech Technology

The extensive application of generative speech technology has profoundly changed scenarios such as dubbing narration, artificial broadcasting, and reading recitation. Based on the current development trend, AIGC technology is expected to empower scenarios such as video and audio dubbing narration, artificial intelligence anchor broadcasting, and audiobook content production, thereby enriching the dissemination patterns of spoken language.

3.1 Video and Audio Dubbing Narration Scenarios

On social media platforms, technological development has made the production process of short videos increasingly simplified. Software like "Jianying" and "Douyin" has even introduced "one-click movie" production methods. Short videos have always been a "battlefield" for major platforms and capital in the Internet world, and generative speech technology can achieve cost reduction and efficiency enhancement in producing short videos [6].

In major sports events, AI intelligent commentary generates commentary content by perceiving the scene and adjusting in time according to the habits of various languages to meet the language requirements. For example, during the Hangzhou Asian Games, the AI commentary system developed by Peking University cooperated with the TV sports channel. The recorded broadcasts of table tennis, taekwondo, diving, and other events called on the multimodal large model capabilities to understand the video content and generated commentary in Chinese, English, Xizang language, Arabic, and other languages to serve people with different language needs.

This is also the first practice and attempt of multimodal generative artificial intelligence in an international large-scale comprehensive sports event.

3.2 Artificial Intelligence Anchor Broadcasting Scenario

Artificial intelligence anchor broadcasting includes virtual hosts (digital people) and artificial intelligence emotional hosts. The application scenarios of the two are distinctly different. Virtual hosts are simulated human images processed by digital technology and interact with the audience through broadcasting, networks, and other communication media [7]. Virtual hosts appeared earlier, and "Ananova," launched by a British network company in 2001, was the world's first virtual host. In addition, Japan launched Yuki [8], China launched Alana, the United States launched Vivian, and South Korea launched Lusia. China's television industry also had a craze for virtual hosts, such as CCTV's "E-mail" and Jiangsu TV's "QQ Miss." With the continuous development of artificial intelligence technology, major radio and television stations have launched their virtual anchors. They are comparable to real people with precise voice synthesis technology, realistic face synthesis technology, and holographic imaging technology. These virtual anchors are not only lifelike in appearance but also perform well. For example, China Central Radio and Television Station launched the ultra-realistic anchor "AI Wangguan," based on financial commentator Wangguan. It deeply imitates Wangguan in voice, mannerisms, and actions. It appears in "Crown Observes the Two Sessions" with the real Wangguan, explaining new policies, conveying multiple important information, and cooperating tacitly. To a certain extent, the launch of virtual anchors is a supplement and expansion of the traditional broadcasting method, making up for the accurate grasp of history, culture, and other backgrounds when people host or broadcast in language. In addition to their application in traditional media, virtual hosts are widely used as receptionists and guides on corporate websites and shopping platforms, providing guidance services when real hosts are not present [9].

In artificial intelligence emotional robots, artificial intelligence digitizes and symbolizes human emotions, and after deep learning, it can recognize, express, and expand human emotions to a certain extent. The key technology of artificial intelligence emotional robots is based on big data, cloud computing, artificial intelligence, deep neural networks, and other intelligent voice technologies, including voice recognition, semantic recognition, semantic analysis, semantic production, semantic synthesis, and voice synthesis [10]. For example, the artificial intelligence companion virtual robot "Xiao Bing" hatched by Microsoft's "Xiao Bing Framework" in 2014, uses full-duplex voice interaction perception (Full et al. Sense) technology to predict what humans are about to say in real-time, generate responses in real-time, and control the rhythm of the conversation, making long-term voice interaction possible. Intelligent hardware devices using this technology no longer require users to say the wake-up word in each round of interaction. They only need to wake up once to easily achieve continuous conversation. Artificial intelligence emotional hosts' emergence and iterative upgrade is currently a research hotspot. It integrates computer science, psychology, vocal science, linguistics, and other disciplines, aiming to create a more emotionally colored voice and expression in the future.

3.3 Audiobook Content Production Scenario

In recent years, the development momentum of China's audiobook industry has been rapid. According to the "2022 China Digital Reading Report," more than 30% (35.5%) of Chinese adults have the habit of listening to books in 2022, an increase of 2.8% compared to the previous year. Audiobook reading has become a major way of reading for Chinese nationals. In the audio field, AIGC mainly lands in three scenarios: voice synthesis, text generation of specific voices, and music arrangement. For example, Microsoft's GhatGPT and AI voice tool VALL-E have higher fidelity than machine-synthesized voices, which will enhance the listening experience of users with audiobooks [11].

In the digital reading field, digital reading companies such as Yuewen, Fan Shu, Ximalaya, and Qi Mao have been transformed and upgraded, highlighting the concept of

"reading +." For example, Tomato Novel is a free reading product launched by Douyin Group in 2019, and the Volcano Voice team supports its AI reading. It now has 40 different voice colors adapted to different themes of novels, such as fantasy, romance, and power struggle. With the support of national policies for audiobook content production, the audiobook market has gradually matured, and audiobooks with high playback volume are mainly concentrated in five types: suspense and thriller, fantasy and fantasy, romance, humor and humor, and children's fiction. For example, "Mi Xiaoquan's School Days (Grades One, Two, and Three)" has an annual playback volume of 3.4 billion times on Qingting FM. The application of generative speech technology in the digital reading field enhances the user's listening experience and serves audiobook platform scene-based research.

In media content authorization, applying TTS technology has greatly improved users' convenience of receiving information. In 2022, the audio client "Yunteng" of China Central Radio and Television Station, with the help of AI voice editing and processing technology, initially formed an industrialized and mass production of news information. It not only outputs content but also far exceeds the efficiency of manual work in terms of quantity. Furthermore, in 2023, "Yunteng" audio introduced AI technology into the mixed editing of digital information and some content review businesses, relying on the early foundation of "Yunteng" to layout in the smart carport, creating an AI radio that is more suitable for individual preferences to meet users' customized and personalized needs. In addition, some website articles or WeChat public account push articles can be directly converted into audio files compatible with TTS technology, which users can listen to while browsing or downloading for offline listening. It also provides more convenient information services for more than 17.8 million visually impaired people in China. This function simplifies how users obtain information and is an important development direction for generative speech technology.

4. Hidden Worries and Digital Environmental Risks of Generative Speech Technology Application

While generative speech technology empowers media platforms, some problems cannot be ignored. First, the AI dubbing generated by the software leads to a low threshold for video generation, and the homogenization of content is serious, with monotonous content and timbres making users feel aesthetically tired. Second, in the process of AI's independent creation, there will also be issues of copyright infringement and privacy leakage, for which academia has already discussed the abuse of AI dubbing behavior. Lastly, under the support of the algorithm's large model, the way users receive information has undergone fundamental changes, and the information cocoon effect makes users change from actively "searching for information" to passively "receiving information," which is not conducive to retaining their deep thinking ability. With the weakening of mainstream media discourse power, the decentralized communication between users is prone to creating public opinion whirlpools.

4.1 Content Generation Threshold Reduction, Homogenization Serious

On the one hand, popular trends lead to the homogenization of content on network video platforms; on the other hand, the low-priced and fast-food culture of short videos also shapes user thinking. Nowadays, short video platforms are filled with many alternative and funny content, coupled with low-priced and easily obtained AI dubbing, and these creative content all try to obtain larger traffic by stimulating users' emotions and desires [12]. Some video creators who use AI dubbing regard the voice as an appendage of the video, which only plays a role in content communication, resulting in many homogenized short video works. In addition, the rapidity of generative AI also makes many video creators have no time to consider the ideological and aesthetic aspects of the work when seizing network hotspots. They only want to harvest hot spot traffic the first time.

In addition, the short-term pursuit of a certain style of timbre by users also leads to the emergence of many homogenized videos. For example, in the current popular AI dubbing, in addition to being distinguished by personality and timbre, there are also some popular roles in film and television dramas, such as "Monkey Brother," "Sire," and "Zhao Wei."

These roles rely on the spread of film and television works and have already formed a thinking set in users' cognition. Combined with specific copywriting, they are prone to produce unexpected fun. However, the proliferation of such video content will also cause the audience's aversion. American anthropologist Hall believes that high-context culture mainly relies on the context for communication, while low-context culture mainly relies on information encoding for information dissemination [1]. The homogenization of video content is a manifestation of low-context culture.

4.2 Existence of Technology Misuse Risks, Re-examination of Communication Ethics

AI dubbing is cost-effective, technically simple, and highly operable, attracting many users. However, we will inevitably encounter "copyright issues when using AI to create video content." First, if AI dubbing uses copyrighted text, music, or video materials for creation, it may infringe on copyright. Second, AI voice synthesis may also infringe on users' voice rights because generative voice synthesis technology requires a large amount of voice library training to clone human voices, and the Internet makes it extremely easy and difficult to regulate the acquisition of voice sources [13]. Unlike the portrait rights protected by legislation, human voices, although highly recognizable, are not protected by law, and only the voices that form works are protected by copyright law. Therefore, there is no clear law to regulate the infringement issues of voice works generated by AIGC technology.

In addition, using AI technology to clone human voices also increases the difficulty of distinguishing the authenticity of information. Assistant Professor Dane Nguyen from the School of Electronic and Data Engineering at the University of Technology in Sydney said, "Artificial intelligence models can recreate voice frequencies based on relatively short voice segments and splice the segments into coherent sentences. For some artificial intelligence models and algorithms, the required time is less than a minute." To obtain improper benefits, it is important to be vigilant against using generative speech technology by voice forgers in certain special fields [14]. For example, there have been reports in the news that criminals use AI to synthesize the voices of victims to extort and blackmail their families, seriously threatening the property rights and right to life of others. In addition, if AI technology is used to clone the voices of opinion leaders with a certain discourse power to publish inappropriate remarks, it will not only form negative online public opinion but also cause harm to the reputation of others. Therefore, AIGC technology is a double-edged sword, and protecting user privacy and data security is particularly important.

4.3 Discourse Power Migration, Creating Public Opinion Whirlpools

Foucault believes that "real power is realized through discourse." The two-way interactivity of online media provides a broad space for user expression, and discourse power is no longer controlled by certain authoritative classes. At the same time, the migration and exchange of the identities of content creators, publishers, and recipients on the platform have created changes in the communication model, which, to a certain extent, dissolves the situation where traditional media monopolizes the discourse power of information dissemination [15]. The widespread application of AIGC technology has intensified the migration of discourse power.

From another perspective, the migration of discourse power has also brought many negative effects. Compared with the communication environment of traditional media, network communication, under the support of AIGC technology, is dissolving mainstream public opinion. Network users spread unverified hot comments in the virtual space anonymously, making it easy to form a public opinion whirlpool [16]. Neil Postman described the risks brought by the misuse of media technology, "The loss of control, overflow, trivialization, and bubble of information make the world difficult to grasp, and people may become slaves to information." Although Postman's worries are overly pessimistic, they remind us of the serious consequences of information being out of control. Therefore, we must be vigilant against the negative effects of the disorderly migration of discourse power and the formation of public opinion whirlpools.

5. Conclusion

Speech, as the material shell of language, is

not only a physical phenomenon but also carries human emotions and plays a role in conveying emotions and meanings in interpersonal communication. As a core application of human-computer interaction, generative speech technology lacks the joys and sorrows of human emotional expression and appears rational but insufficient in sensibility. However, with AI technology's continuous development and maturity, the future will consider ethical and legal issues through technology and algorithm upgrades, protecting user personal privacy and data while achieving personalized voice generation and making generative speech technology a beneficial tool for users' lives and creation rather than a source of negative social public opinion.

References

[1] Li D, Su T. Opportunities and concerns: The great transformation of news productivity is driven by AIGC - Observation and analysis based on the ChatGPT phenomenon. New Media and Fusion Observation, 2023, (04), 15-21.
[2] Di J. Principles of AIGC technology and its application in new media micro-video creation. Applied Mathematics and Nonlinear Sciences, 2024, 9(1), 123-136.
[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. Communications of the ACM, 2020, 63(11), 139-144.
[4] Jiang T, Xu Y, Fu S, et al. Human-intelligence interaction experience research: Injecting new momentum into the development of human-based artificial intelligence. Library Information Knowledge, 2022, 39(4), 43-55.
[5] Li B, Bai Y, Zhan X, et al. Technical characteristics and morphological evolution of artificial intelligence generated content (AIGC). Library Information Knowledge, 2023, 40(01), 66-74.
[6] Feng J. A study on the development path of AIGC empowering short video language. Journalism Research Guide, 2024, 15(06), 1-4.
[7] Lu K, Leng Y. Analysis of user scene experience in the AIGC era - Taking the AIGC application creation studio of Xinhua News Agency's audio department as an example. Journalism Knowledge, 2024, (03), 55-59+95.
[8] Weng J. Research on the impact of intelligent voice technology on the profession and industry of broadcasting and hosting. TV Research, 2017, (12), 57-59.
[9] Zhang W. Research on strategies for improving news editing and creation ability. News Culture Construction, 2024, (04), 151-153.
[10] Han X. Zhou E. Leap in ability and strategic restructuring: A path analysis of generative artificial intelligence driving media deep integration. Chinese Editor, 2024, (02), 29-35.
[11] Luo J. Regulatory landscape and implications of China's artificial intelligence generative content (AIGC) industry. Modern Economics & Management Forum, 2023, 4(4), 64-73.
[12] Gu X, Li J. Research on the application of AI painting technology in short video effects. Silk Screen Printing, 2023, (16), 92-94.
[13] Yu G, Teng W. Developing explainable artificial intelligence: Constructing and governing digital inequality in the AIGC era from a resilience perspective. Contemporary Communication, 2024, (04), 10-14.
[14] Zhou Y, Qin F. Reflections on the application of AIGC in online audiovisual content production - From the perspective of technology philosophy. Young Journalist, 2024, (03), 87-90.
[15] Peng L. How does intelligently generated content affect human cognition and creation? Editor's Friend, 2023, (11), 21-28.
[16] Chesney R, Citron D K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 2019,107, 1753.