Humanities and Social Sciences
2026, Vol. 14, No. 3, pp. 209–214
https://doi.org/10.11648/j.hss.20261403.12

SciencePG
Science Publishing Group

Research Article

# Mutual Shaping: AI-Generated Voice and Meaning-Making in Audiobook Listening

**Yuexi Su***

Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, the United States

## Abstract

This article's primary goal is to examine audiobooks as a context through which the interaction between human and Synthesized Voice generated by AI has become socially meaningful. More broadly, this study approaches audiobooks as a communicative process in which meaning is not transmitted but constructed through interaction between human listeners and technological agents. The new researches indicates that the audiobooks in social-technical context created new interaction between human listeners, making the AI-generated voice socially meaningful. While audiobooks have traditionally been treated as a convenience-based extension of print reading, Recent developments in AIGC have transformed audiobooks into a context in which listeners engage not only with narrative content but also with synthetic voice as a socially interpretable entity. This shift is making AI generated audiobook narration to be fundamentally experiential. Unlike traditional audiobooks, which voice is perceived as an actor. In other words, artificial intelligence-synthesized sound is a symbol of social significance. In the context of audiobooks, it interacts deeply with people, shapes people's cognition as part of people's daily life. In this context, speech synthesis functions as a media technology that participates in an actor network with listeners, whereby human and technological agents reciprocally influence one another and collaboratively produce social significance. This paper makes three key contributions. First, it reconceptualizes audiobooks as sociotechnical interaction sites rather than passive media formats, highlighting the active role of synthetic voice in shaping user experience. Second, it integrates CASA and ANT to provide a multi-level theoretical framework that explains synthetic voice as both a perceived social actor and a networked actor. Third, it proposes a four-stage mechanism model that specifies how social meaning is constructed through interaction with AI-generated narration.

## Keywords

Audiobooks, Synthetic Voice, CASA, ANT, Social Meaning, Human–AI Interaction

## 1. Introduction

Existing research has primarily examined audiobooks in terms of usability, accessibility, and narrative engagement, often treating voice as a functional component of content delivery. Similarly, studies on synthetic voice have largely focused on perception, evaluation, and human–AI interaction at the individual level. However, these approaches tend to overlook how synthetic voice operates as part of a broader sociotechnical system in which meaning is co-constructed through interaction. This paper addresses this gap by shifting the analytical focus from isolated perception to relational processes,

*Correspondence: Yuexi Su (Bettysu@usc.edu)

**Received:** 7 April 2026; **Accepted:** 6 May 2026; **Published:** 15 May 2026

emphasizing how synthetic voice acquires social significance through ongoing interaction within audiobook contexts. Therefore, this paper asks how AI-generated synthetic voices in audiobooks acquire social meaning through interaction with listeners within sociotechnical contexts.

# 2. Literature Review

## 2.1. Audiobooks as a Listening Context

From a communication perspective, audiobooks can be understood not simply as media formats but as sites of mediated interaction where meaning emerges through interpretation and engagement. Audiobooks are typically defined as an electronic book format that is listened to instead of being read in the traditional sense. [1] It recorded readings of written texts intended for listening instead of visual reading. Under the context of studies of digital consumption and technology, audiobooks were given a broader definition of recordings of books or magazines. Audiobooks in the current day have been commonly applied in streaming platforms, audiobook apps, and mobile listening.

Primarily, audiobooks serve as a role of a writing- storage device that includes an audio dimension with phonetic script. [2] Rather than simply an alternation of reading format, audiobooks also function as a distinct media form structured by vocal performance, sound design and auditory temporality. [1] The distinctive traits of audiobooks make an audiobook a book author, a professional actor, an amateur, or a synthetic voice. [3] Given the complex role of audiobook voices, the narrator's voice performance also affects message delivery. An expressive narrator capable of conveying subtle emotional shifts creates an immersive listening experience for users. [1] Therefore, the shift from real humans to AI-generated narrators introduces new questions about how synthetic voices shape listeners' perception, emotional engagement, and interpretation of narrative content.

Traditionally, audiobook voices rely on professional voice actors or author narrations, but the rise of AI allows synthetic voices to become a new option for audiobook designers. [3] The developments of AI's text- to- speech system allow platforms to automatically generate narration immediately. Compared to the traditional narrations, synthetic voices saved a significant amount of time and production costs. The efficiency advantage of AI-generated content allows synthetic voices to be increasingly integrated into audiobook ecosystems. The integration of AIGC-synthesized voices into audiobook platforms further shifts people's perceptions of audiobook voices. [4] When consuming AI-generated narration from an audiobook, users encounter synthetic voices as agents delivering text content. This shift makes the characteristics of the voice itself an important object of analysis, as listeners do not simply process the linguistic content but also interpret the social cues embedded in vocal signals.

## 2.2. Voice as Social Signal

Voice has been proven to be a socially meaningful communicative channel rather than a neutral carrier of linguistic information. In Belin's work Thinking the Voice: Neural correlates of voice perception, human voices are processed by specialized neural mechanisms, forming 'auditory face', that allows us to recognize individuals and emotional states. [5] Similar to how people infer from the traits on face, listeners generate personality impressions from voices. [6] In other words, listeners rapidly form impressions about speakers after hearing minimal vocal input, while also convey emptions and social intention based on the message the listener perceived. [7] This observation indicates that acoustic cues such as pitch, tone, and rhythm function as powerful signals in social judgment. [8]

Beyond formulating social impressions, voice also turns impressions into broader social meaning. According to the research from Nass and Brave, voice as a channel during communication can deliver personality, emotion, identity, and intention, whereas people infer gender, personality traits, and trustworthiness through voice only. [4] In addition, emotional vocal expressions also play a crucial role in the interactive communication between the audience and the voice. When visual cues are absent, individuals rely heavily on vocal signals to infer emotional states, identity, and communicative intent. [9] This demonstrates that voice itself carries social meaning independent of semantic content, which allows listeners to recognize affective cues. In addition to conveying affect, vocal characteristics influence how listeners evaluate speakers. Research shows that pitch, timbre, and speech rhythm significantly affect perceptions of warmth, competence, and trustworthiness. [10]

## 2.3. CASA Framework

Given that voices carry social meaning, questions arise when such vocal cues apply to synthesized voices generated by AI. In fact, the social significance of voice becomes particularly important in the context of synthetic speech technologies. Firstly, do people respond to the voices socially although the voices come from a non-human source. Theoretical perspective has explained this. According the theory of Computers Are Social Actors (CASA), not only do machine voices deliver social messages to audiences, but people also respond socially to machine voices. The Computers Are Social Actors (CASA) paradigm proposes that people respond to media technologies using the same social rules they apply to human interaction. [11] Therefore, when humans interact with computers, people often respond socially to media technologies with human-like cues. The sounds in audiobooks are synthesized by artificial intelligence technology and are a kind of nonhuman source. [12] Rather than consciously deciding to anthropomorphize machines, users often automatically apply social scripts when interacting with technological systems.

The reaction users make to the computer comes from instinctive social habits. This effect has been demonstrated across numerous contexts, including computers, virtual agents, and conversational interfaces. Thus, users tend to attribute personality, politeness, and intentionality to media systems even when they know these systems lack human consciousness.

Though the CASA theory can be applied broadly to multiple media technologies, different modalities function differently. Among these cues, voice is particularly powerful. Considering voice's special characteristic of delivering emotion, identity and intention in its communication channel, voice integrates affective and social information that triggers interpersonal interpretation. Voice is perceived not only as a carrier of content but also as an expression of a speaking subject. Therefore, applying the theory to the context of synthetic voices, voice characteristics function as social cues that evoke audiences' perceptions of personality, emotion and intentionality. [4, 11] Even though the voices were generated by machines, they retain their functions of activating social meanings using the social cues it carries. As a result, listeners would naturally interpret the synthesized voices in social and relational terms. In this sense, machine-generated voices can still become the communicative agent when narrating audiobooks' scripts.

### 2.4. ANT Framework

A related theory known as ANT theory further explains why this phenomenon occurs. The ANT theory refers to Actor- Network theory, which primarily conceptualizes society as a heterogeneous network composed of both human and nonhuman actors, including technologies, artifacts, and infrastructures. [13, 14] The theory argues that society is not only constructed by human beings, but also by networks. The network includes human, technology, media, and texts, which can be concluded in a word: "actors". To conclude, anything that could shape behaviors and perceptions can be defined as actors, and actors will reconstruct societal relationships. Therefore, synthesized voices also serve as the role of an actor. Within this framework, technological objects are not merely instruments used by humans. Instead, they participate in social action by stabilizing relationships, mediating interactions, and shaping communicative practices. ANT, therefore, shifts analytical attention from isolated individuals to relational networks in which agency emerges through connections among actors.

From an ANT standpoint, actors do not possess intrinsic agency independent of context. Agency is relational and distributed across networks. [13] Technologies can thus function as what Latour terms "quasi-objects", which can be defined as entities that gain social force through their participation in networks of practice. In this context, technologies mediate how social meaning is constructed during communication. As the theory has proven technologies participate in social action, there are three ways actors act. The first one is mediation, as summarized by Latour, is the "intermediary transports meaning without transformation, while a mediator transforms, translates, and modifies what it carries." [13] In other words, technologies mediate how communication is structured by shaping how information produces its own meaning and how the meaning being interpreted. The second one is stabilization. The technologies also stabilized the communication pattern. According to Law's work Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity, networks do not exist naturally; instead, the network was stabilized through constant repeat on communication patterns. [15] Taking synthesized voices as an example, when the same type of AI narration is repeatedly using, audiences would form a stabilized mode of listening, emotional expectation and narrative interpretation. Therefore, this also makes ANT theory is relevant for AI-generated voices, since AI-generated voices' perceived agency emerges through interaction with listeners, platform infrastructures, and narrative contexts. Lastly, translation also provides a crucial lens for understanding how nonhuman actors participate in social process. In the context of ANT theory, translation can be understood as actors within a network transforming and reconfiguring relations among other actors. [14] The translation process helps connections to be established, and helps interaction to be reorganized, such that identities and relationships of actors are continuously negotiated within the network.

### 2.5. Integration

Integrating CASA and ANT allows for a multi-level explanation of how synthetic voice acquires social meaning. Building on this integrated theoretical perspective, a process-oriented model is needed to explain how social meaning is gradually constructed through interaction. While CASA accounts for the cognitive and perceptual processes through which listeners respond socially to human-like vocal cues, ANT extends this perspective by situating synthetic voice within a broader network of relations involving listeners, platforms, and narrative structures. This integration highlights that the social significance of synthetic voice is neither purely psychological nor purely technological. Instead, it emerges through the dynamic interaction between human perception and sociotechnical networks, where meaning is continuously negotiated and stabilized.

## 3. Theoretical Model / Mechanism

To further specify how synthetic voice becomes socially meaningful, this paper proposes a four-stage mechanism of social meaning construction. Rather than treating user response as a single process, this model conceptualizes interaction with synthetic voice as a layered and cumulative progression, in which perception, interpretation, and behavior are gradually shaped through ongoing engagement. These stages are sequential and cumulative, with each stage building upon the previous one to shape increasingly complex forms of

social meaning.

### 3.1. Affective Perception

The process begins with the perception of vocal cues and initial affective responses. Regarding technology's influence on humans, this process is evident in four aspects. The first aspect concerns affective response. When listening to audiobooks, audiences rely heavily on vocal cues and make affective and evaluative judgments based on vocal characteristics, which in turn shape varying levels of perceived intimacy and trust toward the voice. [7, 16] This process begins with the perception of vocal cues and serves as a medium that carries emotional and interpersonal messages. On the other hand, listeners are still able to infer affective states even with the absence of visual cues, indicating that voice is a powerful channel in interpreting communication messages. During the perception of voice, listeners infer affective state and personal attitude toward the voice, which further translates into evaluation. The evaluation of voice reshaped the audience's attitude to voice as well. The voice under the audience's personal evaluation is no longer experienced as a neutral technology, but begins to take on qualities of a socially present entity. As mentioned earlier, the CASA theory indicates that audiences treat computer-generated media exhibiting human-like traits as human agents. Once the voice successfully formed social entity, meaning that it is socially becoming more similar to a real human. Thus, there will also be an interpersonal connection being built between the listener and the voice. This engagement also enhances immersion, since there is a stronger emotional resonance with the voice deepened in listener's perception, which further raise listeners' involvement in the audiobook content and strength the experiential connection with the narrative. [17] The voice, overall, triggers a multiple-stage process, which is from simple perception to affect, then shaping one's evaluation of the message, and lastly enhances the audience's immersive experience when listening to audiobooks.

### 3.2. Interpretive Engagement

Beyond simply precepting the voice, listeners integrate their interpretation of narrative content with their imagination of the voice, through which sound functions as a crucial mediating channel for meaning-making that creates imagination constructed experience. The voice actively shapes the voice's implication by guiding attention, pacing interpretation and framing emotional tone. [1] In this sense, Listeners integrate their interpretation of narrative content with their own imagination on the synthesized voice, making voices function as a crucial mediating channel of meaning perception. In this process, audiences interact with the voice, engage in forms of self-disclosure, and experience emotional belonging and value affirmation. These imaginative engagements create a form of perceived interaction. During the interaction, listeners respond to the voice as if it were addressing them, leading to a sense of dialogic involvement. The involvement also evokes audiences a sense of belonging since listeners resonate the voice with their own emotional states, which may further become value affirmation. [18, 19] The second stage, therefore, brings out another social function of synthetic voice, which is a site where meaning is constructed by the interaction between auditory experience and the listener's self-perception. During the stage, the synthetic voice completed the process of transforming from simple human ai interaction to reshape self-value.

### 3.3. Persona Construction

The third stage concerns the construction of imagined voice personae. In this stage, listeners construct imagined voice personae (or avatars) based on acoustic cues such as pitch and emotional valence. In audiobooks, acoustic cues are the primary source of social and affective information for listeners. The information allows listeners to infer personality traits, emotional states and communicative intentions, thereby forming a coherent mental representation of the voice as an entity. [10] Through ongoing interaction with these imagined entities, they may develop affective attachment and even dependence, gradually incorporating the voice into their everyday routines and experiential lifeworld. This process aligns with concepts in anthropomorphism. Anthropomorphism research likewise indicates that humans have a cognitive bias toward attributing humanlike qualities to nonhuman entities, particularly when those entities display communicative behaviors such as speech. [20] As a result, imagined voice personas may be perceived as social agents because audiences are willing to attribute humanlike qualities to the synthesized voice. As voice personas are repeatedly accepted as social agents, the exposure and interaction between humans and AI allow listeners to develop affective attachment to these imagined personas, such as emotional comfort and reliance. The attachment may further led listeners to incorporate the voice into their everyday life routines, and treat the voice as a meaningful presence that is closely aligned with listeners' real life. In this way, synthetic voice becomes an imagined social identity that shapes audiences' ongoing emotional engagement and lived experience.

### 3.4. Behavioral and Relational Influence

The fourth aspect suggests that vocal sound can actively shape listener behavior, influencing responses, decisions, and patterns of engagement. This influence operates through the persuasive qualities of vocal cues, which shape how information is processed and responded to by listeners. [21] Experimental evidence suggests that vocal characteristics alone can significantly influence user decisions and compliance, even when message content remains constant. [22] In such cases, listeners rely on peripheral cues embedded in the voice rather than the semantic content of the message,

leading to variations in compliance and decision-making. [21] This highlights the persuasive potential of voice as a communicative channel and suggests that AI-generated narration may actively shape user experience rather than functioning as a passive delivery mechanism. As a result, the vocal characteristics directly influence users' responses, such as user's behavioral intention and compliance. These effects may further shape the broader pattern of users' engagement with synthesized voice in audiobooks, as they influence how frequently listeners interact with audiobooks and how they incorporate voice-based media into their daily life routine. Thus, another function that synthetic voice performs is to act as a persuasive agent that shapes users' behavior and engagement actively. [11] To conclude, the voice not only operates a force within audiences' listening experience, but also influences how listeners act upon it.

### 3.5. Summary

Taken together, these four stages suggest that when technology indeed exhibits socially meaningful signals that operates as a layered and cumulative process. These signals performed in voices are likely to be perceived as a social actor. The role of AI-generated voices in audiobooks extends beyond perception to include interaction and mutual shaping, through which listeners interpret these voices and construct them as socially meaningful symbols. Beginning with affective response, listeners first interpret vocal cues as emotional signals, which shape their evaluation and perceived intimacy toward the synthesized voice. The evaluation then leads to engagement that shapes the audience and voice's relationship, which the audience recognizes as a real social agent. Building on this, synthetic voice is being reconstructed into an entity with real social meaning that audiences could resonate with. In this sense, synthetic voice should be understood not merely as a communicative medium but as an evolving sociotechnical entity whose meaning is continuously constructed and negotiated across contexts.

## 4. Discussion

The proposed framework also carries broader implications beyond theoretical explanation. Beyond theoretical implications, the framework of AIGC audiobooks also has practical significance. As AI-generated voices are increasingly utilized in audiobooks and digital media platforms, understanding how voice design influences social perception and user engagement becomes crucial for both the producer and audience. Designers and platforms must recognize the change in AIGC content performance. First, synthetic voice is not neutral but actively shapes user experience, emotional response, and behavioral outcomes. At the same time, this also raises important ethical and design considerations. If synthetic voices are capable of shaping users' emotions, perceptions, and behaviors, their deployment cannot be treated as a purely technical decision. AI as an actor is capable of actively influencing individuals, since they are being recognized as an agent that is socially present. This involves questions of transparency, manipulation, and user autonomy. For instance, highly optimized synthetic voices designed to maximize engagement or persuasion may blur the boundary between assistance and influence, raising concerns about whether users are fully aware of how their responses are being shaped. This is particularly relevant in contexts such as education, mental health support, and personalized media consumption, where emotional and cognitive effects are more pronounced. Therefore, future design practices should not only focus on improving realism and efficiency but also consider how to maintain ethical boundaries, ensure user awareness, and support meaningful rather than manipulative interaction with AI-generated voices.

## 5. Conclusion

In summary, this paper has examined how AI-generated synthetic voices in audiobooks function as socially meaningful entities through interaction. As a non-human sound source, an audiobook originally has no social significance. However, in the context of audio books, combined with the content of the reading materials, it can have an impact on the audience, which is manifested in the establishment of a sense of intimacy, emotional absorption, and integration into daily life practices, etc. This is the domestication of technology by humans. The new perspective discussed in this paper challenges the assumption that AI-generated narration is merely a functional substitute for human voice. Instead, synthetic voices should be understood as active participants in communication, shaping how meaning is produced, interpreted, and experienced. In the personal space, the media technology that originally belonged to the public space is regulated and acquires social significance. On the other hand, technology also makes adjustments in response to human demands to better serve humanity. This framework also opens new directions for future research, particularly in examining how variations in voice design, platform affordances, and cultural contexts influence the social perception of synthetic voices. As AI-generated narration becomes increasingly widespread, understanding its social implications is critical.

## Author Contributions

**Yuexi Su:** Formal Analysis, Methodology, Writing – original draft, Writing – review & editing

## Conflicts of Interest

The author declares no conflicts of interest.

## References


[1] Have, I., Pedersen, B. S. Digital audiobooks: New media, users, and experiences. Routledge; 2021.

[2] Rubery, M. Audiobooks, literature, and sound studies. Routledge; 2011.

[3] Nguyen, H. V., Phan, T. T., Nguyen, H., Tran, V. T., Nguyen, N. Understanding audiobook apps' consumption values and their implications for promoting audiobooks in Vietnam. Publishing Research Quarterly. 2023, 39(1), 61–78. https://doi.org/10.1007/s12109-022-09934-w

[4] Nass, C., Brave, S. Wired for speech: How voice activates and advances the human–computer relationship. MIT Press; 2005.

[5] Belin, P., Fecteau, S., Bédard, C. Thinking the voice: Neural correlates of voice perception. Trends in Cognitive Sciences. 2004, 8(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

[6] Todorov, A., Olivola, C. Y., Dotsch, R., Mende-Siedlecki, P. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annual Review of Psychology. 2015, 66, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

[7] Scherer, K. R. Vocal communication of emotion: A review of research paradigms. Speech Communication. 2003, 40(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

[8] Baus, C., Couto, B., Escera, C., Costa, A. Forming social impressions from voices. Scientific Reports. 2019, 9, 1–9. https://doi.org/10.1038/s41598-018-36518-6

[9] Neves, L., et al. Neural decoding of emotional prosody. Cerebral Cortex. 2021, 33(3), 709–721. https://doi.org/10.1093/cercor/bhab090

[10] McAleer, P., Todorov, A., Belin, P. How do you say "hello"? Personality impressions from brief novel voices. PLOS ONE. 2014, 9(3), e90779. https://doi.org/10.1371/journal.pone.0090779

[11] Reeves, B., Nass, C. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press; 1996.

[12] Gambino, A., Fox, J., Ratan, R. Building a stronger CASA: Extending the computers are social actors paradigm. Human–Machine Communication. 2020, 1, 71–85. https://doi.org/10.30658/hmc.1.5

[13] Latour, B. Reassembling the social: An introduction to actor-network theory. Oxford University Press; 2005.

[14] Callon, M. Some elements of a sociology of translation. In: Law, J., Ed. Power, action and belief. Routledge; 1986, pp. 196–223.

[15] Law, J. Actor network theory and material semiotics. In: The new Blackwell companion to social theory. Wiley-Blackwell; 2009, pp. 141–158.

[16] Juslin, P. N., Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin. 2003, 129(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

[17] Kuzmičová, A. Audiobooks and the imagination. In: The Routledge companion to literature and science. Routledge; 2016, pp. 375–386.

[18] Silverstone, R. Why study the media? Sage; 1999.

[19] Couldry, N. Theorising media as practice. Social Semiotics. 2004, 14(2), 115–132. https://doi.org/10.1080/1035033042000238295

[20] Epley, N., Waytz, A., Cacioppo, J. T. On seeing human: A three-factor theory of anthropomorphism. Psychological Review. 2007, 114(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

[21] Petty, R. E., Cacioppo, J. T. Communication and persuasion: Central and peripheral routes to attitude change. Springer; 1986.

[22] Jiang, Z., et al. Manipulative power of voice characteristics in human–AI interaction. Proceedings of the ACM on Human-Computer Interaction. 2024.