Toward a Third-Kind Voice for Conversational Agents in an Era of Blurring Boundaries Between Machine and Human Sounds

Jeesun Oh
sun.oh@kaist.ac.kr
Industrial Design, KAIST
Daejeon, Republic of Korea

Hyeonjeong Im
imhyeonjeong@kaist.ac.kr
Industrial Design, KAIST
Daejeon, Republic of Korea

Sangsu Lee
sangsu.lee@kaist.ac.kr
Industrial Design, KAIST
Daejeon, Republic of Korea

ABSTRACT

The voice of widely used conversational agents (CAs) is standardized to be highly intelligible, yet it still sounds machine-generated due to its artificial qualities. With advancements in deep neural networks, voice synthesis technology has become nearly indistinguishable from a real person. The voice enables users to discern the speakers' identities and significantly impacts user perception, particularly in voice-only interactions. While more natural, human-sounding voices are generally preferred, their use in CAs raises potential ethical dilemmas, such as eliciting unwanted social responses or confusing the nature of the speaker. In this evolving landscape, it is necessary to understand the voice characteristics from multiple facets of voice design for CAs. Therefore, our study examines the voice characteristics of both artificial-sounding and human-sounding voices. Then, we propose a 'third-kind' of voice that considers the characteristics of each voice type. This discussion contributes to the debate on the future direction of voice design in the field of Conversational User Interface research.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**.

KEYWORDS

Voice user interface; Voice interaction; Voice assistant; Voice-based conversational agent; Speech synthesis; Human-sounding voice; Artificial-sounding voice; Transparency; Efficiency; Naturalness

ACM Reference Format:
Jeesun Oh, Hyeonjeong Im, and Sangsu Lee. 2024. Toward a Third-Kind Voice for Conversational Agents in an Era of Blurring Boundaries Between Machine and Human Sounds. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3640794.3665880

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665880

1 INTRODUCTION

The early stages of voice synthesis were based on the concatenative synthesis, which re-combined recorded unit selections, resulting in consistent and hyper-articulated speech [17]. This approach has now achieved a top-notch level of intelligibility and is commonly employed in widely adopted conversational agents (CA) such as Alexa, Siri, and Google Assistant [11]. These CAs typically have a standardized, broadcaster-like voice with a neutral accent and no disfluencies — what we call 'artificial-sounding'. People can clearly understand these voices, but also easily distinguish them from human voices because of their excessive clarity. In recent years, with the advent of deep neural network synthesis [53], unique human voice inflections that were not present in traditional machine-generated voices are now being incorporated. It can produce sounds that are almost indistinguishable from humans, more than just human-like — what we call 'human-sounding voices' [25]. Human-sounding voices can even mimic human disfluencies (e.g. filled pauses), the unique accent, emotions, or human noises, such as laughter, yawning, and coughing [33–35, 45, 54, 55, 61, 67]. Tech companies have surprisingly replicated celebrity voices, including those of John Legend (Google) [8], Samuel L. Jackson (Amazon) [3], and many others[1, 72].

Numerous studies have indicated that human-sounding voices are generally perceived as more likable and positive compared to artificial-sounding voices [7, 11, 29, 37, 50, 51, 60]. Users tend to perceive computers as social actors, regardless of their mechanical nature [41]. They also prefer voices that share similar characteristics with their own (e.g., personality [38], ethnicity [42]) according to the similarity-attraction theory [39]. However, simply pursuing a human-sounding voice for CAs without proper consideration can potentially raise ethical concerns. It might mislead users through mindless emotional reactions [41], confuse them about the identification of the conversation partner [57], and further inadvertently deceive them, making it difficult to discern whether the source is a machine or a human [32]. Voices can convey more than just linguistic content. Beyond words, a voice can represent phonetic features such as prosody, intonation, and speech rate and it also portrays personal traits like gender, age, and regional accents [39, 42]. Given these voice characteristics, the voice integrated into the CA plays an important role in enabling users to identify the nature of the speaker and significantly influences user perception, especially in a voice-only context.

In an era where the distinction between machine and human sounds is increasingly blurred, it is crucial to seek the direction of voice design considering multifaceted aspects. Blindly choosing CAs' voices to give a product a next-generation feel would not be sufficient. Our study first delves into three considerations of CA voice design: transparency, efficiency, and naturalness. Then, we examine the characteristics that distinguish between artificial-sounding and human-sounding voices to understand which voice characteristics make it challenging for users to differentiate CAs from humans and which features make them feel natural and familiar. Based on this understanding, we discuss the voice design direction and strategies for CAs. Consequently, we suggest a 'third-kind' of voice for CAs that ensures transparency by clearly conveying their mechanical nature and efficiency for transactional purposes, while also maintaining natural, human-sounding characteristics that users find familiar and comfortable. This discussion contributes to the growing debate on future voice design direction that should be pursued for CAs in the Conversational User Interface (CUI) research field.

2 THREE CONSIDERATIONS FOR CONVERSATIONAL AGENT VOICE DESIGN

To explore the direction of CA voice design from multiple aspects, we examine three critical considerations—transparency, efficiency, and naturalness—and explain why they are important in voice interactions.

2.1 Transparency in Addressing Ethical Concerns

Even though a CA with a human-sounding voice allows users to perceive the conversant as more familiar and facilitates natural communication, particularly in voice-only interactions, it poses an ethical dilemma by making it difficult for users to identify whether the speaker is a human or a machine. Aylett [4] emphasizes that it is difficult for users to detect mimicry in synthetic speech, describing it as "less of a valley and more of an abyss" in terms of the uncanny valley.

Firstly, the issue arising from the lack of transparency in the voice can confuse users about the identity of their conversational partner. In the study conducted by Shanka [58], it was reported that an elderly lady called her cellular provider to make changes to her plan. The customer service representative, without clearly identifying whether it was human or AI, behaved similarly to a human operator, simulating the sound of typing on a keyboard whenever the user spoke. This created the impression that the agent was looking up information. This led the old lady to feel confused and uncertain about interacting with a person or an AI agent. Once she determined that she was dealing with an AI, in her state of confusion, she found herself wishing only to be connected to a human.

To address this issue, merely having a CA with a human-sounding voice disclose its nature does not resolve. Even though users are explicitly informed that they are interacting with a CA using a human-sounding voice, they might still become confused and suspicious. They may wonder whether they heard correctly or even change their minds midway through, believing they are interacting with a human due to the difficulty of distinguishing voice mimicry. For example, Google Duplex openly introduced itself as 'Google's automated booking service.' However, during the conversation with the CA, the restaurant manager remained doubtful about its identity and indirectly inquired by asking a curveball question, "Are there any kids?" to ensure that it was a person [13].

In addition, even if users are aware that a human-sounding voice is from a CA, they may become distracted out of surprise at the advancement of technology or curiosity to test the system's capabilities by asking bait questions that divert them from the main task. While users' amazement and curiosity will soon fade as they get used to the technology, it can still lead to inefficiencies in system operations like customer service in the early stages of adoption. Moreover, due to the fleeting nature of its voice, background noises, or a poor connection, users might miss the CA's self-introduction. No matter how overtly a CA reveals its identity, there is the potential to unintentionally deceive or confuse users due to its indistinguishable human-sounding voice, particularly in a voice-only context. Therefore, resolving such concerns about transparency is not as simple as just revealing a CA's identity. Similar to how users can identify personal traits (e.g., gender, age, and region) from vocal characteristics alone, it is necessary to explore the voice characteristics that inherently represent the nature of a CA.

2.2 Efficiency for Transactional Purposes

Human spoken conversation can broadly be classified as either transactional or social interaction [9]. Although widely used CAs like Apple's Siri and Amazon's Alexa are considered to combine both transactional and social interactions [49], users primarily engage in voice conversation with CAs for transactional purposes [16, 56]. To effectively support users in accomplishing transactional tasks, the clarity of speech synthesis must be fundamentally considered in voice design, because a clear and high-quality voice reduces users' cognitive load [24]. Moreover, voice designers emphasize 'beyond-human characteristics,' such as machine-like speed, to ensure transactional tasks are performed efficiently. Therefore, efficiency in voice design is important for supporting users successfully complete transactional tasks through CAs.

2.3 Naturalness for Seamless Interactions

Over the years, research has examined the impact of computer-synthesized voice compared to human speech. People generally prefer the human voice [7, 11, 29, 37, 50, 51, 60]. They also tend to feel more comfortable interacting with human-like voices of CAs [43, 71]. Schroeder et al. [52] revealed that the key difference between computer speech synthesis and human speech lies in the naturalistic variance of paralinguistic cues. Accordingly, the naturalness of a voice can help users feel comfortable and foster more seamless interactions. However, this also means that the naturalness of the voice could potentially mislead users' social responses. In Moore's [36] study, when users called travel agents, customers were more likely to engage in lengthy social exchanges (an 83% increase) with normal human voices than with a robotic-sounding voice, resulting in inefficient task completion. Another study by Wang et al. [68] reported that human voices generated more customer complaints than AI systems in call centers. Therefore, naturalness is significant in voice design not only for facilitating comfortable and smooth interactions but also for regulating engaging social interactions.

3 ARTIFICIAL-SOUNDING VOICE AND HUMAN-SOUNDING VOICE

We operationally define artificial-sounding and human-sounding voices and compare them to understand their respective voice characteristics.

3.1 Artificial-Sounding Voice: Transparently Recognizable as a Conversational Agent

The term 'artificial-sounding voice' is used differently across various studies: non-human voice [36], robotic voice [36], the default voice style [4], one-fits-all voice [11], standard synthetic voices [20]. However, they are being interpreted in a similar notion. Since the widespread adoption of voice-based CAs (e.g., Amazon's Alexa, Apple's Siri, and Google Assistant) by global companies from 2015, many smart devices equipped with CAs have adopted a standardized default voice. Cambre [12] mentioned it as a "one-fit-all voice" which has a clear, female-sounding, polite, and playful voice. Aylett [4] also describes "the default voice style" as having a newsreader style, being clear and warm but unemotional. This speech synthetic technology has advanced from its past low-quality robotic-sounding to highly intelligible sound [11, 21]. The early stages of voice synthesis were based on the concatenative synthesis, which re-combined recorded unit selections, resulting in hyper-articulated speech, but also included prosodic peculiarities [17]. Subsequent advancements and the integration of new models like a parametric synthesis and neural network, have facilitated the seamless co-articulatory overlap (i.e., more connected speech), thereby achieving the upper limit of artificial-sounding speech. As a result, it can deliver information with a high level of clarity, much like professional TV news reporters, yet still maintains a discernible difference from the casual speech of real humans. The voices also feature standard pronunciation and intonation, consistent speed and volume, and relatively uniform stable tone and pitch, all without any disfluency [12, 21, 52]. Based on this, we operationally define 'artificial-sounding' as standardized, emotionally neutral, articulate, broadcaster-like, and delivered in a relatively friendly manner.

3.2 Human-Sounding Voice: Almost Indistinguishable from Human Voice

The advancement of deep neural network synthesis has enabled speech synthesis to reach a level of realism that was incredibly difficult to discern from actual human voices. However, the interpretation of 'human-sounding voice' varies within the HCI discipline. This is to be expected as human speech is a vast domain, with extensive study in related fields like Linguistics and Phonetics. Google Duplex and subsequent studies used the term "natural-sounding" which is associated with the filled pauses and synthetic latency [30, 45]. In a recent research by Do [20], Neural TTS was studied, yielding higher fidelity (i.e., increasingly smooth and natural) in intonation-based deep neural network models. Likewise in the tech industry, the extent to which human-sounding voice technology is being progressed is diverse. Neural speech synthesis models, such as Deepmind's Wavenet [23], Google's Tacotron 2 [69, 70], Baidu's Deep Voice 3 [6], and more models [46] can generate synthetic voices with fluent and more natural prosody and characteristic accents reflecting individuality, ethnicity, and regional dialects. Not only can they capture 'one of a kind' accents, but models developed by Microsoft's Vall-e [35, 67], Elevenlabs [55], D-ID [33], and Neosapience's Typecast [54] also incorporate human emotions in speech. Other models, such as Meta's Generative Spoken Language Model (GSLM) [34], and Suno's Bark AI [61], can even mimic sounds uniquely human noises, for example, laughter, yawning, coughing, or mouth clicks. Driven by advancements in these technologies, we also define operationally 'human-sounding' as voices that intricately mimic the unique voice quality of human speech, making them nearly indistinguishable from the voices of real people.

3.3 Voice Characteristics Between Artificial-Sounding Voice and Human-Sounding Voice

To comprehend the vocal characteristics of 'artificial-sounding' and 'human-sounding' voices, we conducted a literature review and classified seven features, as shown in Table 1.

3.3.1 *Prosody.* The dynamic prosody, including rhythm, pitch, and intonation, adds expressiveness to speech. Schroeder [52] found that paralinguistic cues are instrumental in generating a more mindful human voice. The study revealed that an authentic human voice exhibits higher pitch, dynamic intonation, and more frequent pauses compared to a computer-generated voice. On the other hand, artificial-sounding voices tend to have relatively stable, calm intonations and lack expressivity [27, 52].

3.3.2 *Speech Rate.* Speech rate refers to how quickly or slowly words are spoken. Koiso et al. [28] discovered that human speech tends to start relatively slowly at the beginning and accelerates towards the end of the discourse. They also found that the human speech rate is related to the information structure in dialogues. In contrast to human-sounding voices exhibiting irregular speech rates, artificial-sounding voices display a consistent speech rate [52].

3.3.3 *Disfluency.* Disfluencies indicate natural hesitations and imperfections in human speech, such as pauses, filled pauses, filler words, repetitions for self-corrections, context-dependent omissions, and verbosity [31]. The study also noted that pauses are very common in human speech, with an average of one pause occurring every 49 words. Google Duplex simulates filled pauses, like 'hmm. . . ' and 'uh. . . ' to create a natural-sounding effect [30, 45]. Unlike human-sounding speech, which contains disfluencies, artificial-sounding voices are very intelligible and free from disfluencies [11, 21].

3.3.4 *Response Latency.* The response latency is the short or long response time of a conversation partner. Leviathan [30] reported that people expect an immediate response after a simple 'hello?' and more latency in response to complex sentences [45]. In dialogues, longer latency could represent analytical reasoning [45, 52]. Such latency is distinguishable from a long delay in computers. On the

Table 1: Classification of voice characteristics between artificial-sounding and human-sounding voices.

Features	Artificial-Sounding Voice	Human-Sounding Voice
Prosody	Stable intonation; sounds calm [52]	Dynamic intonation; sounds cheery [21, 52]
Speech rate	Consistent [52]	Inconsistent (Irregular) [28, 52]
Disfluency	Clarity without disfluency [11, 21]	More pauses and filled pauses (e.g., uhm…, uh…) [45]
Response Latency	Consistent according to the system processing time: Guided to respond within 2 seconds [59]	Varies according to context: Respond to instantly or slowly [30, 45]
Accent	Standardized accent [11]	Characterized accent [45, 62]
Emotions	Unemotional [4, 21, 27]	Emotional [21, 63]
Noise	Semantic-free noises such as beeps, squeaks, and clicks from robots in moives [5]	Human noises such as breathing, yawning, coughing, and sneezing from the body's organs [34, 61]

other hand, the latency of artificial-sounding voices is dependent on system processing. When latency lasts too long (more than two seconds) or occurs at inappropriate times, users feel awkward and perceive it as a long delay while interacting with the computer [59].

3.3.5 Accent. The accent captures the subtle nuances of human speech, including variations in pronunciation, dialectal intonation, and stress patterns. Human-sounding voices are enriched by diverse sociolects (e.g., African American Vernacular English), social classes (e.g., posh accent), and regional and national accents [35, 72]. Sutton [62] introduced sociophonetics, which explores social qualities such as geography and social class associated with voice output [1, 55]. However, artificial-sounding voices typically use the most standardized accent, similar to the pronunciation of newsreaders.

3.3.6 Emotions. Many studies have identified four dimensions of emotions (activation, valence, potency, and intensity) and have shown that emotions such as happiness, sadness, fear, disgust, and anger can be conveyed through vocal expressions [48]. Cowen [18] also found that vocal bursts can convey 24 emotions, including awe, pain, relief, sympathy, and more. These emotions can be expressed in human-sounding voices [53, 63]. However, artificial-sounding voices, which generally sound pleasant, have a consistent speech rate and relatively stable prosody, limiting their ability to express varied emotions and often resulting in a voice that sounds dull and emotionless [4, 21, 27].

3.3.7 Noises. The human voice produces various noises due to bodily functions, including yawning, snoring, hiccupping, coughing, sneezing, breathing, and tongue clicking. Meta's Generative Spoken Language Model (GSLM) is capable of mimicking sounds such as yawning and mouth clicks [34]. Additionally, Suno's Bark AI can imitate vocalizations like gasps and throat clearing [61]. In contrast, artificial-sounding voices typically do not generate these types of sounds produced by the human body. Instead, they can produce semantic-free utterances often heard in movies or media portraying robots, such as the 'beep-boop' sounds from R2D2 or the hovering and clanking sounds from Wall-E [5].

4 VOICE DESIGN DIRECTION AND STRATEGIES

4.1 At the Crossroads of Voice Choice: The Third-Kind of Voice

We discuss the direction of voice design for CAs regarding which voice characteristics are useful and necessary in both artificial and human-sounding voices, in terms of transparency, efficiency, and naturalness.

4.1.1 Transparency and Efficiency in Artificial-sounding Voice. To ensure transparency in voice interactions with CAs, it is important to carefully design unique voice traits that could mislead users into thinking they are interacting with humans. CAs should use standardized accents rather than characterized accents that are specific to regions, countries, or social classes. Even though Sutton et al. [62] emphasized the importance of sociophonetics in voice design for diversity and individualization, these characterized accents might confuse users about the system's origin or activate users' social stereotypes of cultural and historical backgrounds [44]. Similarly, human noises should be avoided in voice interactions with users. Instead, mechanical sounds that can be immediately recognized as coming from CAs could be utilized. In addition, to allow users to perform transactional functions with CAs efficiently, the intelligibility of the voice is essential. Clarity should not be compromised in an effort to achieve a human-sounding nuance. Thus, the voice of CAs should not include disfluencies such as filled pauses and communicate clearly and articulately to promote effective voice interactions for transactional operations with users.

4.1.2 Naturalness in Human-Sounding Voice. Just as important as transparency and efficiency, communicating naturally is significant as well. Responsive and dynamic voice attributes such as prosody, speech rate, and response latency should be adopted for natural and comfortable voice interactions with CAs. These features enable users to converse easily and freely, allowing for smooth turn-taking in conversations without hesitation or restraint. Although emotions are inherently human, the emotional tone of a voice could be subtly adjusted to resonate empathetically with the user's emotions, thereby promoting engaging voice interactions. In Kim's study [27], voice interaction designers indicated that expressing emotions through CA voices with cheerful or sorrowful tones, according to the user's sentiment, can foster a positive relationship. Similarly, Chin [14] found that agents who displayed empathy were most effective in managing users' verbal abuse. However, social CAs that feign emotion and empathy could lead to superficial and inauthentic relationships [64, 65]; hence, emotion in voice design requires a meticulous approach.

4.1.3 Third-Kind Voice: Transparent, Efficient, and Natural. Rather than choosing between artificial-sounding and human-sounding voices, we suggest a third-kind of voice that pursues three aspects. In summary, to ensure transparency of the speaker's identity as manifested in the voice, CAs should avoid misrepresenting uniquely human voice qualities such as regional accents or human noises. Additionally, they should exclude disfluencies and maintain clear articulation for efficient communication. Moreover, to facilitate natural interactions, it is important to provide responsive and sophisticated prosody, speech rate, and response latency. However, the emotional tone in the voice should be delicately adjusted to foster empathy aligned with the user's emotions and manner.

Although previous studies have attempted to generate distinct voices by adding unique mechanical sounds [5] or creating genderless voices [19], there has been a notable gap in efforts to develop a third-kind voice that is neither entirely human-sounding nor artificial-sounding, while considering transparency, efficiency, and naturalness. We believe that it is important to pioneer a third-kind voice for CAs that considers the aforementioned three aspects. Echoing our proposal for a third-kind of voice that blends artificial-sounding and human-sounding voice characteristics, similar implications have emerged from other studies. Clark [16] suggested treating CAs as "a new genre of conversation with its own rules, norms, and expectations." Kim [26] argues for a machine-like approach where CAs should have task-oriented purposes as machines. Additionally, Kim [27] asserts that a natural voice user interface should selectively adopt certain aspects of human naturalness rather than emulating every aspect of natural human conversation. In the following sections, we will describe some design strategies for a third-kind voice for CAs, along with existing examples.

4.2 Design Strategies for the Third-Kind Voice

While maintaining naturalness in the voice of CAs, we suggest design strategies that enhance transparency and efficiency.

4.2.1 Layering Mechanical Sounds. We suggest incorporating mechanical or futuristic sounds, such as bleeps, jingles, or electronic static (whir), into the background or intermittently throughout a third-kind of voice to help users recognize their interaction with a system. As mentioned in section 3.3.7, Aylett et al. [5] investigated mixing speech synthesis with mechanical sounds (i.e., semantic-free utterances (SFUs)) such as beep-boop, squeaks, and clang-clanks from Wall-E, BB-8 and R2D2, drawing on the rich history of robotic sounds in films like Star Wars. Cambre [11] also indicated that voice interfaces could use non-human features like "earcons" – brief audio clips signaling activity or status in screen readers. In addition, such distinctive sounds can be utilized to manage response latency. Instead of using the human metaphor like 'umm…', specific digital sounds can signify the system's processing for complex tasks, akin to a visual loading icon. For example, distinctive semantic-free sounds can be utilized during loading times. Such design approaches not only alert users that a response is being generated, thereby reducing perceived wait time [59], but also reinforce the understanding that they are communicating with a system. Therefore, distinct mechanical sounds can be used intermittently within a speech or as loading sounds during response latency.

4.2.2 Imparting Otherworld Voice Textures. We propose to directly integrate complex modulations of 'otherworldly' voice textures into a third-kind voice. For example, the AI protagonist's voice in the movie 'Tau' exhibits dynamic and sophisticated paralinguistic cues but possesses a digital voice texture that is unmistakably non-human, making it familiar yet obviously mechanical. But since assigning excessive mechanical voice texture could be perceived as a threat, the use of an emotionless or monotonously rigid voice, similar to the voice of 'Auto,' the main antagonist in the movie 'Wall-E,' should be reconsidered cautiously. Such voices can trigger users' stereotypes that filmmakers often depict robots and machines as threats to humankind [66]. In a similar vein, previous studies have attempted to design gender-neutral voices. Apple has also released a gender-neutral voice named 'Quinn' [47]. However, attempts to design gender-neutral voices have not been favored by users [19]. Although gender-neutral voices are not typically preferred, we advocate the idea that CAs should have gender-neutral voices because gender is a predominant human-kind voice quality, which might inadvertently provoke users' gender stereotypes [40]. So, maintaining natural prosody while employing a gender-neutral or digitally distorted voice texture can serve as a strategic approach to clearly signify its machine origin.

4.2.3 Enabling Manipulation of Voice Attributes. Not only does it imbue distinct characteristics to the voice itself, but it also facilitates interactions that allow for the manipulation of voice attributes. The voice interface could be designed to give users free control over their interactions, enabling them to barge in, speak quickly, skip steps, and resume conversations even if interrupted in the middle. Amazon's Alexa has been equipped with a feature that allows users to control the speech rate by saying 'speak faster' or 'speak slower' [2]. Choi et al. [15] also found that individuals with visual impairments not only perceive the CA as a social actor but also expect to have the ability to control speech rates, similar to their control with screen readers. In Kim's study [27], voice interaction designers stated that a natural voice user interface should feature machine-specific capabilities, such as allowing CAs to speak in a blazing-fast manner to accomplish transactional tasks efficiently. Accordingly, in managing user interruptions during voice interactions [10, 22], the voice user interface should allow users to easily interrupt in order to manipulate voice attributes. Thus, the voice user interface could be designed to enable users to freely interrupt and control features like speech speed, replay, and skipping, enhancing both efficiency and transparency.

5 CONCLUSION

In the era of blurring lines between machine and human sound, our study compared artificial and human-sounding voices of CAs. We provided an operational definition and analyzed the voice characteristics of both artificial and human-sounding voices. This enables design practitioners and researchers to better comprehend each voice characteristic and to contemplate which features should be endowed to the voice of CAs. Drawing from these voice characteristics, we discussed the CA's voice design direction considering three aspects of transparency, efficiency, and naturalness. While maintaining the naturalness of the voice, we suggested adding mechanical sounds and providing a distinct texture to the voice itself to enhance transparency. We also proposed manipulating vocal features, such as speech rate, to improve efficiency. Although pioneering studies have shown a low user preference for these voice transparency approaches [5, 19] the pursuit of a balance between transparency, efficiency, and natural voice characteristics must continue. Moving forward, a 'third-kind voice' — one that users not only prefer but also find beneficial — should be further explored and refined. This exploration will drive the discovery of a voice for CAs that aligns with the intrinsic nature of the machine, transcending both current artificial and human-sounding voices.

REFERENCES

[1] Lawrence Abrams. 2023. Bing Chat has a secret 'Celebrity' mode to impersonate celebrities. https://www.bleepingcomputer.com/news/microsoft/bing-chat-has-a-secret-celebrity-mode-to-impersonate-celebrities/. Accessed: 15 August 2023.

[2] Amazon. 2019. Alexa, speak slower. https://www.aboutamazon.com/news/devices/alexa-speak-slower. Accessed: 15 August 2023.

[3] Amazon. 2020. Samuel L. Jackson celebrity voice for Alexa gets an update. https://www.amazon.science/latest-news/samuel-l-jackson-celebrity-voice-for-alexa-gets-an-update. Accessed: 15 August 2023.

[4] Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You Have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3290607.3310422

[5] Matthew P. Aylett, Yolanda Vazquez-Alvarez, and Skaiste Butkute. 2020. Creating Robot Personality: Effects of Mixing Speech and Semantic Free Utterances. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 110–112. https://doi.org/10.1145/3371382.3378330

[6] Baidu Research. 2017. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. http://research.baidu.com/Blog/index-view?id=91. Accessed: 15 August 2023.

[7] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. 2018. The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. In *Interspeech*. https://api.semanticscholar.org/CorpusID:52191344

[8] Manuel Bronstein. 2019. Hey Google, talk like a Legend. https://blog.google/products/assistant/talk-like-a-legend/. Accessed: 15 August 2023.

[9] Gillian Brown and George Yule. 1983. *Discourse analysis*. Cambridge university press.

[10] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 911–920.

[11] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376789

[12] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 223 (nov 2019), 19 pages. https://doi.org/10.1145/3359325

[13] Brian X. Chen and Cade Metz. 2019. Google's Duplex Uses A.I. to Mimic Humans (Sometimes). https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html. Accessed: 15 August 2023.

[14] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[15] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376569

[16] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[17] Michelle Cohn and Georgia Zellou. 2020. Perception of Concatenative vs. Neural Text-To-Speech (TTS): Differences in Intelligibility in Noise and Language Attitudes. In *Proceedings of Interspeech*. 1733–1737. https://doi.org/10.21437/Interspeech.2020-1336

[18] Alan S. Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2019. Mapping 24 emotions conveyed by brief human vocalization. *The American psychologist* (2019). https://api.semanticscholar.org/CorpusID:58563174

[19] Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. https://doi.org/10.1145/3544548.3581281

[20] Tiffany D. Do, Ryan P. McMahan, and Pamela J. Wisniewski. 2022. A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 424, 11 pages. https://doi.org/10.1145/3491102.3517564

[21] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[22] Patrick Gebhard, Tanja Schneeberger, Gregor Mehlmann, Tobias Baur, and Elisabeth André. 2019. Designing the Impression of Social Agents' Real-time Interruption Handling. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 19–21. https://doi.org/10.1145/3308532.3329435

[23] Google DeepMind. 2016. WaveNet: A generative model for raw audio. https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio. Accessed: 15 August 2023.

[24] A. Govender and S. King. 2018. Measuring the cognitive load of synthetic speech using a dual task paradigm. In *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018*. International Speech Communication Association, 2843–2847. https://doi.org/10.21437/Interspeech.2018-1199 Conference code: 139961.

[25] Xuedong Huang. 2018. Microsoft's new neural text-to-speech service helps machines speak like people. https://azure.microsoft.com/en-us/blog/microsoft-s-new-neural-text-to-speech-service-helps-machines-speak-like-people/. Accessed: 15 August 2023.

[26] Hyeji Kim, Inchan Jung, and Youn-kyung Lim. 2022. Understanding the Negative Aspects of User Experience in Human-Likeness of Voice-Based Conversational Agents. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1418–1427. https://doi.org/10.1145/3532106.3533528

[27] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 242, 13 pages. https://doi.org/10.1145/3411764.3445579

[28] Hanae Koiso, Atsushi Shimojima, and Yasuhiro Katagiri. 1998. Collaborative Signaling of Informational Structures by Dynamic Speech Rate. *Language and Speech* 41, 3-4 (1998), 323–350. https://doi.org/10.1177/002383099804100405

[29] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics* 14 (2020). https://doi.org/10.3389/fnbot.2020.593732

[30] Yaniv Leviathan. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html. Accessed: 15 August 2023.

[31] Robin Lickley. 1994. Detecting disfluency in spontaneous speech. *The University of Edinburgh* (01 1994). http://hdl.handle.net/1842/21358

[32] Natasha Lomas. 2018. Duplex shows Google failing at ethical and creative AI design. https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/. Accessed: 15 August 2023.

[33] Aisha Malik. 2023. D-ID's new web app gives a face and voice to OpenAI's ChatGPT. https://techcrunch.com/2023/03/07/d-ids-new-web-app-gives-a-face-and-voice-to-openais-chatgpt/. Accessed: 15 August 2023.

[34] Meta AI. 2021. Textless NLP: Generating expressive speech from raw audio. https://ai.facebook.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/. Accessed: 15 August 2023.

[35] Microsoft. [n. d.]. Vall-E. https://www.microsoft.com/en-us/research/project/vall-e-x/. Accessed: 15 August 2023.

[36] Roger K Moore. 2017. Appropriate voices for artefacts: some key insights. In *1st International workshop on vocal interactivity in-and-between humans, animals and robots*.

[37] John W Mullennix, Steven E Stern, Stephen J Wilson, and Corrie lynn Dyson. 2003. Social perception of male and female computer synthesized speech. *Computers in Human Behavior* 19, 4 (2003), 407–424. https://doi.org/10.1016/S0747-5632(02)00081-X

[38] Clifford Nass and Kwan Min Lee. 2000. Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 329–336. https://doi.org/10.1145/332040.332452

[39] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153 arXiv:https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/0022-4537.00153

[40] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1997.tb00275.x

[41] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[42] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.

[43] Andreea Niculescu, Betsy Dijk, Anton Nijholt, Haizhou Li, and Sl See. 2013. Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *International Journal of Social Robotics* 5 (04 2013), 171–191. https://doi.org/10.1007/s12369-012-0171-x

[44] Andreea Niculescu, George M. White, See Swee Lan, Ratna Utari Waloejo, and Yoko Kawaguchi. 2008. Impact of English Regional Accents on User Acceptance of Voice User Interfaces. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges* (Lund, Sweden) *(NordiCHI '08)*. Association for Computing Machinery, New York, NY, USA, 523–526. https://doi.org/10.1145/1463160.1463235

[45] Daniel E. O'Leary. 2019. GOOGLE'S Duplex: Pretending to Be Human. *Int. J. Intell. Syst. Account. Financ. Manage.* 26, 1 (mar 2019), 46–53. https://doi.org/10.1002/isaf.1443

[46] OpenAI. 2024. Navigating the Challenges and Opportunities of Synthetic Voices. https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices. Accessed: 15 August 2023.

[47] Sarah Perez. 2022. Siri gains a new gender-neutral voice option in latest iOS update. https://techcrunch.com/2022/02/24/siri-gains-a-new-gender-neutral-voice-option-in-latest-ios-update/. Accessed: 15 August 2023.

[48] Patrik Juslin Petri Laukka and Roberto Bresin. 2005. A dimensional approach to vocal expression of emotion. *Cognition and Emotion* 19, 5 (2005), 633–653. https://doi.org/10.1080/02699930441000445 arXiv:https://doi.org/10.1080/02699930441000445

[49] Silvia Quarteroni. 2018. Natural language processing for industrial applications. *Spektrum* 41, 2018 (2018), 105.

[50] Emma Rodero and Ignacio Lucas. 2023. Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society* 25, 7 (2023), 1746–1764. https://doi.org/10.1177/14614448211024142 arXiv:https://doi.org/10.1177/14614448211024142

[51] Simon Schreibelmayr and Martina Mara. 2022. Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology* 13 (2022). https://doi.org/10.3389/fpsyg.2022.787499

[52] Juliana Schroeder, Michael Kardas, and Nicholas Epley. 2017. The Humanizing Voice: Speech Reveals, and Text Conceals, a More Thoughtful Mind in the Midst of Disagreement. *Psychological Science* 28, 12 (2017), 1745–1762. https://doi.org/10.1177/0956797617713798 arXiv:https://doi.org/10.1177/0956797617713798 PMID: 29068763.

[53] Dagmar M. Schuller and Björn W. Schuller. 2021. A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice. *Emotion Review* 13, 1 (2021), 44–50. https://doi.org/10.1177/1754073919898526 arXiv:https://doi.org/10.1177/1754073919898526

[54] Eric Hal Schwartz. 2022. New Neosapience Tool Synthesizes Any Text into Emotion for Virtual Actor Speeches – Exclusive. https://voicebot.ai/2022/09/14/new-neosapience-tool-synthesizes-any-text-into-emotion-for-virtual-actor-speeches-exclusive/. Accessed: 15 August 2023.

[55] Eric Hal Schwartz. 2023. Synthetic Speech Startup ElevenLabs Raises $2M for AI Voices With Context-Relevant Emotion. https://voicebot.ai/2023/01/23/synthetic-speech-startup-elevenlabs-raises-2m-for-ai-voices-with-context-relevant-emotion/. Accessed: 15 August 2023.

[56] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[57] Daniel B. Shank, Christopher Graves, Alexander Gott, Patrick Gamez, and Sophia Rodriguez. 2019. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior* 98 (2019), 256–266. https://doi.org/10.1016/j.chb.2019.04.001

[58] Daniel B. Shank, Christopher Graves, Alexander Gott, Patrick Gamez, and Sophia Rodriguez. 2019. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior* 98 (2019), 256–266. https://doi.org/10.1016/j.chb.2019.04.001

[59] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How quickly should a communication robot respond? Delaying strategies and habituation effects. *International Journal of Social Robotics* 1 (2009), 141–155.

[60] Steven E Stern, John W Mullennix, Corrie-lynn Dyson, and Stephen J Wilson. 1999. The persuasiveness of synthetic speech versus human speech. *Human Factors* 41, 4 (1999), 588–595.

[61] Suno. 2023. Bark. https://github.com/suno-ai/bark.

[62] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300833

[63] Elevenlabs Team. 2022. The first AI that can laugh. https://elevenlabs.io/blog/the_first_ai_that_can_laugh/. Accessed: 15 August 2023.

[64] Sherry Turkle. 2017. Why these friendly robots can't be good friends to our kids. https://www.washingtonpost.com/outlook/why-these-friendly-robots-cant-be-good-friends-to-our-kids/2017/12/07/bce1eaea-d54f-11e7-b62d-d9345ced896d_story.html. Accessed: 15 August 2023.

[65] Sherry Turkle, Cynthia Breazeal, Olivia Dasté, and Brian Scassellati. 2006. Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication* 120 (2006).

[66] Eric Vanman and Arvid Kappas. 2019. "Danger, Will Robinson!" The challenges of social robots for intergroup relations. *Social and Personality Psychology Compass* 13 (07 2019). https://doi.org/10.1111/spc3.12489

[67] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111 [cs.CL]

[68] Lingli Wang, Ni Huang, Yili Hong, Luning Liu, Xunhua Guo, and Guoqing Chen. 2023. Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management* 32, 4 (2023), 1002–1018. https://doi.org/10.1111/poms.13953 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/poms.13953

[69] Yuxuan Wang and RJ Skerry-Ryan. 2018. Expressive Speech Synthesis with Tacotron. https://ai.googleblog.com/2018/03/expressive-speech-synthesis-with.html. Accessed: 15 August 2023.

[70] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[71] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

[72] Cliff Weitzman. 2022. Most famous voice actors. https://speechify.com/blog/most-famous-voice-actors/. Accessed: 15 August 2023.