Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3715336.3735763

RESEARCH-ARTICLE

"Hello, This Is a Voice Assistant Calling": When a Human Voice Calls Claiming to Be a Machine on an Ordinary Day

JEESUN OH, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

YUNJAE CHOI, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

SANG-SU LEE, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Open Access Support provided by:

Korea Advanced Institute of Science and Technology

PDF Download
3715336.3735763.pdf
27 December 2025
Total Citations: 0
Total Downloads: 1645

Published: 05 July 2025

Citation in BibTeX format

DIS '25: Designing Interactive Systems Conference
July 5 - 9, 2025
Madeira, Portugal

Conference Sponsors:
SIGCHI

“Hello, This Is a Voice Assistant Calling”: When a Human Voice Calls Claiming to Be a Machine on an Ordinary Day

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

Yunjae Choi
Industrial Design
KAIST
Daejeon, Republic of Korea
choi.yj@kaist.ac.kr

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

Abstract

With the advent of neural networks, it has become possible to generate synthetic voices that are nearly indistinguishable from real human speech (i.e., human-sounding voice). In contrast, earlier voice assistants used voices that were instantly recognizable as machine-generated, owing to their standardized, consistent, and highly intelligible qualities (i.e., artificial-sounding voice). Although people tend to prefer human-like voices, adopting human-sounding voices in voice assistants raises ethical concerns related to confusion or unintentional deception, particularly in voice-only contexts, even when their identity as systems is explicitly disclosed. To explore the voice design direction for future voice assistants, we examined how participants perceived and interacted when they were unexpectedly confronted with either an artificial-sounding or a human-sounding voice, both of which clearly identified themselves as voice assistants during an everyday phone call. Our findings reveal participants’ experiences and conversational behaviors in each voice condition. Furthermore, we discuss how the voices of voice assistants should be designed and propose design implications that emphasize transparency and responsiveness in voice design.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Voice Interaction, Voice Assistant Call Agent, Voice-Only Context, Speech Synthesis, Human-Sounding Voice, Human-Like Voice, Artificial-Sounding Voice, Wizard of Oz Method

ACM Reference Format:
Jeesun Oh, Yunjae Choi, and Sangsu Lee. 2025. “Hello, This Is a Voice Assistant Calling”: When a Human Voice Calls Claiming to Be a Machine on an Ordinary Day. In *Designing Interactive Systems Conference (DIS ’25), July 05–09, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715336.3735763

This work is licensed under a Creative Commons Attribution 4.0 International License.
DIS ’25, Funchal, Portugal
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1485-6/25/07
https://doi.org/10.1145/3715336.3735763

1 INTRODUCTION

Early models of voice assistants, like Apple’s Siri, Amazon’s Alexa, and Google Assistant, typically used highly intelligible, standardized, broadcast-like female voices with neutral accents and no disfluencies [10, 19]. These voices exhibited consistent, excessive clarity and hyper-articulation, making them distinctly different from human voices. Due to these characteristics, they are immediately recognizable as machine-generated and we refer to such synthetic voices as *artificial-sounding voices*. In recent years, voice synthesis technology has advanced through the incorporation of neural networks. This allows for the production of speech surpassing mere human-like qualities, capturing unique vocal inflections that were absent in traditional artificial-sounding voices. For example, speech synthesis models feature voice characteristics such as filled pauses (e.g., hmm, huh)—as seen in Google Duplex, a Google Assistant booking system [34, 50]—as well as accents reflecting personality, regional dialects, or non-native pronunciations [40, 68]. Building on this trajectory, as of 2024, ChatGPT’s voice mode [48] and Google’s NotebookLM voice overview [67] produce voices that sound strikingly human. However, paralinguistic cues (e.g., nuanced intonation, subtle turn-taking coordination, variations in speech rate, and context-sensitive pause patterns) that responsively adapt to users and context are not yet fully available due to occasional delays and mismatches in real-time turn-taking. Nevertheless, these capabilities continue to advance rapidly. We refer to voices that are almost indistinguishable from human speech, not only in acoustic quality but also in interactional responsiveness, as *human-sounding voices*. Conversations with such human-sounding voice assistants are expected to become feasible in the not-so-distant future.

Previous studies have shown that users, when interacting with human voices in a system, tend to feel more comfortable [46], find them more likable [4, 9, 32, 53, 54], and develop greater trust [72]. However, little is known about the specific benefits and concerns that may arise in conversations with such voices, especially in voice-only contexts. Potential concerns may include user confusion about who (or what) they are interacting with [59], or misinterpretation of social and emotional cues [44]. In an attempt to address these concerns, Google Duplex explicitly introduced itself as “Google’s automated booking service.” Even so, during real-world testing, a restaurant manager remained skeptical about whether the speaker was truly a system [11]. Although a voice assistant clearly and transparently identifies itself as a system, potential concerns may still persist because its human-sounding voice significantly influences user perception. Additionally, user interactions are often shaped by whom they believe and perceive they are communicating with [35, 60, 62]. Given this importance, our research question, positing

a speculative near-future scenario in which voice assistants become almost indistinguishable from humans in conversation, examines how users perceive and interact with them when their identity as a system is explicitly disclosed. To gain richer insights, we also revisit artificial-sounding voices—despite being relatively less likable—to consider whether they still remain a valuable option for addressing ethical concerns in voice design. Therefore, our study aims to empirically explore and compare how users perceive and interact with both an artificial-sounding (immediately recognizable as machine-generated) and a human-sounding (nearly indistinguishable from a human) voice assistant in a voice-only context, where each self-identifies as a voice assistant.

To achieve this, we adopted a Wizard of Oz method to schedule the interview time for our study through real phone calls, in a voice-only context. First, we placed phone calls using either an artificial-sounding or a human-sounding voice, assigned according to a between-subjects design, to expose users unexpectedly and capture their genuine first-time encounters in a real-world, everyday context. Based on multiple pilot tests, we tweaked the Wizard of Oz method to implement a human-sounding voice setup that would be nearly indistinguishable in conversation, in terms of both acoustic quality and vocal responsiveness. In this setup, a human speaker (wizard) delivered the voice, inevitably sounding like a real human while explicitly identifying as a voice assistant. We chose this approach not to reflect the current capabilities of voice technology, but to explore speculative voice interaction scenarios that may emerge in the foreseeable future. Then, we conducted a debriefing interview, which had been scheduled through the aforementioned phone experiment, to examine participants' experiences and the underlying reasons behind their interactions in each voice condition. Drawing on the observation data and debriefing interviews, two researchers carried out both conversation analysis [27, 52] and thematic analysis [8].

Our findings present seven themes, each with related subthemes, describing how participants perceived and interacted with each voice condition, including detailed conversational behaviors and their underlying thoughts and feelings. Most participants interacting with the human-sounding voice condition experienced confusion and disbelief, even when the system was explicitly disclosed. On the other hand, artificial-sounding voices, while less engaging, alleviated social burdens. Building on these findings, we discuss how voice assistants' voices should be designed in a voice-only context: they should clearly reveal their identity through voice alone, foster natural and rich interactions, and relieve social burdens while maintaining empathy. Furthermore, we weave these discussions together to propose voice design implications that strive for both transparency and responsiveness. Our study contributes by providing timely and necessary insights intended to inspire voice interaction designers and researchers to craft the voices of voice assistants—soon to become a pervasive part of daily life. This work also adds to ongoing discussions that are increasingly important to the HCI community, in an era where the line between human and synthetic voices continues to blur.

2 RELATED WORKS

2.1 Artificial-Sounding Voice and Human-Sounding Voice

We compared two key concepts—*artificial-sounding voice* and *human-sounding voice*—and provided operational definitions and relevant background for each to establish a shared understanding.

2.1.1 Artificial-Sounding Voice: Transparently Recognizable as a Machine. The term artificial-sounding voice is used differently across various studies: non-human voice [41], robotic voice [41], the default voice style [1], one-fits-all voice of early voice assistants [9], standard synthetic voices [19], and mindless voices [55]. However, they can be interpreted as variations of a similar concept. From the widespread adoption of initial voice assistants (e.g., Amazon's Alexa, Apple's Siri, and Google Assistant) by global tech companies in 2015, many voice assistants have incorporated a standardized default voice. Cambre and Kulkarni [10] described the one-fit-all voice of early generation voice assistants as clear, female-sounding, polite, and playful. Similarly, Aylett et al. [1] characterized the default voice style of voice assistants at that time as having a newsreader style—clear, warm, but unemotional. Voice synthetic technology has advanced from its past low-quality robotic-sounding to highly intelligible sound [9, 20]. In its early stages, voice technology relied on concatenative synthesis, which recombined recorded speech units. While this approach produced hyper-articulated speech, it also introduced prosodic peculiarities [15]. Later advancements, such as parametric synthesis and neural networks, enabled seamless co-articulatory overlap (i.e., smoother and more connected speech), pushing the boundaries of artificial-sounding speech. These developments have resulted in voices that deliver information with exceptional clarity—akin to professional TV news reporters—yet still differ from the casual, spontaneous speech of real humans, making them sound less human and still artificial. These voices are characterized by standard pronunciation, consistent speed and volume, stable tone and pitch, and an absence of disfluencies [10, 20, 55].

In our study, the term *artificial-sounding voice* is operationally defined as a voice with a standardized accent, consistent speed, emotionally neutral tone, articulate delivery, and a broadcaster-like voice, as widely used in early voice assistants—immediately recognizable as non-human and machine-generated.

2.1.2 Human-Sounding Voice: Nearly Indistinguishable from Human Voice. The advancement of deep neural networks has enabled speech synthesis technology to reach human-level naturalness. Within the HCI discipline, the extent of human-likeness in voice assistants' human-sounding voices varies to some degree. Google Duplex and subsequent studies referred to such voices as "natural-sounding," incorporating features like filled pauses and regional accents (e.g., Irish English) [34, 50]. In research by Do et al., Neural text-to-speech was studied, demonstrating higher fidelity (i.e., increasingly smooth and natural prosody) in intonation-based deep neural network models [19]. Additionally, the tech industry has seen diverse progress in human-sounding voice technology. Neural speech synthesis models, such as DeepMind's Wavenet [25], Google's Tacotron 2 [70, 71], Baidu's Deep Voice 3 [3], and OpenAI's voice model in 2024 [51], can generate synthetic voices with

fluent, natural prosody and distinctive accents that reflect individuality, ethnicity, and regional dialects. The models developed by Microsoft's Vall-E [40, 68], ElevenLabs [57], D-ID [37], and Neosapience's Typecast [56] can now not only replicate unique, 'one-of-a-kind' accents but also integrate human emotions into speech synthesis. Other models like Meta's Generative Spoken Language Model [39] and Suno's Bark AI [63] extend these capabilities to mimic uniquely human noises produced by body organs, such as laughter, yawning, coughing, breathing, and mouth clicks. Even though synthetic voices may sound remarkably human, they still fall just short of fully capturing the paralinguistic cues of human conversation—issues such as processing latency and awkward turn-taking mismatches persist—based on the capabilities of commercial voice interaction technologies as of 2024 (e.g., OpenAI [48]; Google [67]). In particular, paralinguistic cues in speech have been shown to significantly contribute to making voices sound more human [50, 55]. Nonetheless, driven by recent advancements, voice technology is now on the verge of mimicking adaptive and responsive paralinguistic dynamics, such as intonation, speech rate, turn-taking timing, and pauses—ultimately enabling lifelike conversations.

In this work, we have operationally defined a *human-sounding voice* not as one that merely sounds like human speech, but as one that also incorporates paralinguistic dynamics that finely adjust in response to the user and context, enabling interactions that are nearly indistinguishable from those with real humans. Under this definition, while human-sounding voices may encompass a wide range of vocal characteristics, our study intentionally excludes elements such as emotional expression or distinct human noises, as the task involved simple transactional phone calls.

2.2 Comparing Artificial-Sounding Voice and Human/Human-Like Voice

2.2.1 Comparing Artificial-Sounding Voices and Human Voices in a Lab-Based Setting. Previous studies have compared artificial-sounding voices and real human voices in various contexts in the human-computer interaction (HCI) field. Early studies by Stern et al. [61] and Mullennix et al. [42] both found that human speech was generally perceived more favorably than computer-synthesized voices presented via audio tapes. Doyle et al. [20] carried out a qualitative study comparing the voice of smart speakers (Alexa and Siri) and human speech. While both types of voice were found to be equally easy to understand, users perceived the computer-synthesized voice as colder and less expressive compared to human speech, which is described by a cheerful and dynamic tone. Cambre et al. [9] compared 18 commercially available computer-synthesized voices and three human voices in a long-form context like audiobooks. While some of the computer-synthesized voices were perceived as having high clarity similar to human speech, their voice quality (e.g., monotone, naturalness, emotion, and personality) fell short of that of human speech. Kühne et al. [32] compared human voices with two different TTS-synthesized voices and found that human voices were preferred over synthesized voices. They argued that voices with higher human-likeness effectively reduced eeriness and increased likability. Rodero and Lucas [53] had participants listen to audiobooks narrated by human voices and a synthetic voice (Amazon Alexa) and supported the Human Emotional Intimacy Effect, demonstrating that human voices outperform synthetic voices in emotional connection, engagement, and information retention.

2.2.2 Comparing Artificial-Sounding Voices and Human Voices in the Real-World. Some studies have examined both artificial-sounding voices and actual human voices in real-world situations. Moore [41] collected real telephone conversations from a travel planning service and compared interactions between a human voice (telephone operator) and a robotic-sounding voice (produced via a voice changer used by a human speaker). The findings revealed that human callers interacting with the robotic-sounding voice were less likely to engage in lengthy social exchanges, resulting in an 83% reduction in responses. Another study by Wang et al. [69] reported that AI systems in call centers effectively reduced customer complaints. Although these studies contribute to our understanding of how interactions vary between artificial and real human voices in real-world contexts, they still lack an in-depth grasp of user experiences and the reasons why users exhibit limited voice interaction.

2.2.3 Comparing Artificial-Sounding Voices and Human-like Voices. Moving beyond the comparison of artificial-sounding voices to human voices, the latest study compared artificial-sounding voices and deep neural network-based speech synthesis in a controlled lab environment. A recent study by Do et al. [19] investigated the effect of neural text-to-speech (TTS) on user perception in the context of a persuasive virtual human, comparing standard TTS, neural TTS, and human speech. Neural TTS was perceived as less trustworthy than human speech, but no significant difference was found with standard TTS, indicating that higher fidelity synthetic speech is not necessarily more favorable [19]. Earlier researchers [13, 19] also emphasized the importance of re-evaluating earlier research in light of modern speech synthesis technology as it continues to improve. Our study positions itself within the related work landscape and builds on previous research by examining two voice conditions. While the artificial-sounding voice reflects modern technology, the human-sounding voice is based on a near-future perspective and may be revisited as a feasible, real-world technology.

3 METHOD

3.1 WoZ Study on Real-World Phone Calls

To compare users' perceptions and interactions with a voice assistant using both artificial-sounding and human-sounding voices, we used the Wizard of Oz (WoZ) method in a real-world phone call setting, a voice-only context (refer to Figure 1). The WoZ method is widely adopted in the speech-based HCI field [22]. A human operator (the wizard) simulates the interactive behavior of a system, making participants believe they are actually engaging with a functioning system [5, 23]. We regard this method as a useful and effective approach for exploring the two voice conditions while minimizing the risk of conversational processing errors that could disrupt the user experience. Exposing users to such errors through the human voice was not the goal of this study. It is also particularly well-suited to the human-sounding voice, envisioned as a near-future voice interaction that is almost indistinguishable from that of a real human. In addition, in an effort to collect vivid and authentic user experiences, we focused on users' first encounters with

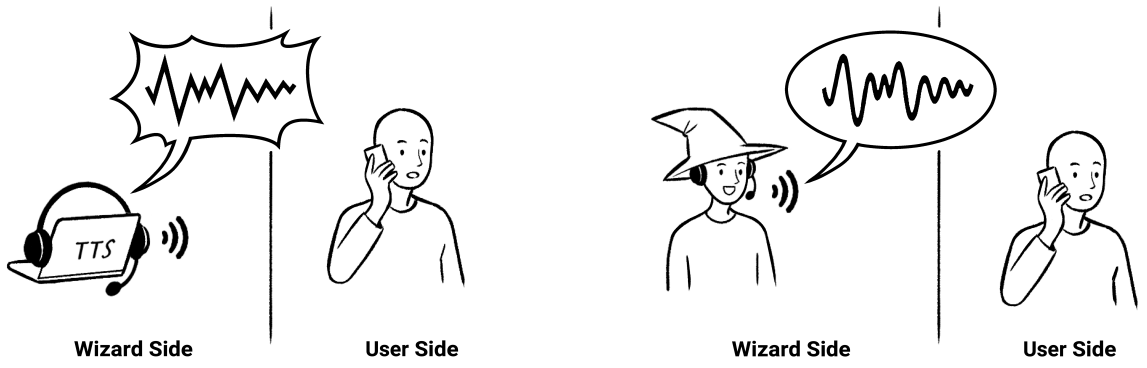

Figure 1: Illustration of our study set in a real-world phone call context. We used a between-subjects design to compare voice conditions. The artificial-sounding voice was produced through text-to-speech, while the human-sounding voice was achieved by a human wizard with voice guidance. Both conditions rigidly followed the dialogue flow and script.

the voice condition in their daily routines. The task involved scheduling the actual interview times for this study through ordinary phone calls. The study was reviewed and approved by the university's Institutional Review Board (IRB). The task of scheduling through phone calls was chosen because it is a very common and familiar task, making it easy for anyone to perform. A voice-only setting requires users to rely solely on their voices, making it ideal for exploring user experiences related to voice-based conditions without the distraction of other elements—such as the visual and physical presence of voice assistants. The phone call environment is also an essential context, as it offers various future applications for AI-based conversational agents.

3.2 Artificial-Sounding Voice Setup

We set up the artificial-sounding voice condition to be used with the WoZ method for our study. The scripts were converted into an artificial-sounding text-to-speech (TTS), and the researcher (wizard operator) simulated responses from a preset list following the dialogue flow. This technique of having the wizard manage pre-recorded prompts is often used in speech-based HCI research [5, 7, 66]. To select the artificial-sounding voice, we assessed five TTS applications that support Korean, as the study was conducted in South Korea with Korean participants: Amazon Polly, Adobe Audition, Google Cloud TTS, Kakao TTS, and Naver Clova Dubbing. After reviewing eight voices from five TTS applications, we opted for the female Ara voice from Naver Clova Dubbing TTS. This voice best represents the artificial-sounding voice operationally defined in Section 2.1.1, exemplifying modernity with consistency, neutrality, intelligibility, and standing out as one of the most advanced options for Korean speech synthesis. The selection of a female voice was made because it is typically used for voice assistants, with no gender-specific reason. The phone calls were relayed using the Skype app on a MacBook Pro, ensuring that the previously prepared TTS audio output was transmitted through the phone call. We used pre-recorded prompts in the simulation and chose not to concurrently employ real-time TTS. This was to guarantee consistent conversational capability and range and to prevent potential delays from the wizard's typing time and the risk of misspellings.

3.3 Human-Sounding Voice Setup

To establish human-sounding voices that are nearly indistinguishable from real human voices in interaction, we determined that having a human directly speak is the most effective and appropriate way to achieve this, much like how generative synthesized voices are produced. Several previous studies have also used a human wizard to directly speak, simulating computer-generated conditions [21, 24, 33, 38, 41]. To set up a truly natural human-sounding voice condition, we determined that the first author—who has prior experience conducting experiments with artificial-sounding voice conditions and has a comprehensive understanding of the human-sounding voice as operationally defined in Section 2.1.2—was the most suitable candidate to act as the wizard. In addition, vocal characteristics of the human wizard, such as gender, age, and accent, were well-aligned with those of the artificial-sounding voice used in the experiment, with the only difference being the voice conditions (i.e., the independent variables) under comparison. The human wizard was a woman with an adult female voice, neither too young nor too old (28 years old), and a standard Korean accent from being born and raised in Seoul. Her accent had inherent personal characteristics but did not contain any regional accents.

Since the human wizard spoke, it was inevitable that instinctive and responsive paralinguistic features were present in the human-sounding voice condition. To ensure a comparable experience across both conditions, several guidelines were carefully followed. First, the human wizard strictly adhered to the dialogue flow and script during her utterances. The script was thoroughly practiced and rehearsed to deliver the most consistent speech performance possible. Second, to maintain consistency, excessive or sudden changes in loudness or speed were avoided; disfluencies—except for filled pauses indicated in the script—were excluded; emotional expression was minimized; and human noises such as laughing, yawning, or coughing were omitted. In all sessions, she made no noticeable reading mistakes (e.g., stumbling over words or conversational hiccups), nor did she significantly deviate from the above guidelines. This was verified by two researchers through a careful review of the recorded data.

3.3.1 Manipulation Check. After the experiment, we conducted a manipulation check to determine whether participants in the human-sounding voice condition were exposed to the wizard's utterances within acceptable dynamic boundaries. For each session, we measured three voice attributes—response time, speaking rate, and pitch—from the wizard's utterances using Praat software. A one-way analysis of variance (ANOVA) was then performed, as it is appropriate for comparing means across multiple independent groups. In this analysis, each participant's session was treated as an independent group (N = 12). The results (response time: $p = 0.49$, speaking rate: $p = 0.85$, pitch: $p = 0.09$) indicate that there were no significant differences in the wizard's utterances across participants.

3.3.2 Pilot Study: Reasons for Not Using Pre-Recording. In our pilot studies before the experiment, we recruited two voice actors to record all scripts using actual human voices, similar to setting up an artificial-sounding voice condition. However, although the recorded human voices preserved human-level acoustic fidelity, they lacked particularly reflexive and subtle turn-taking timing. When this absence of responsive paralinguistic cues failed to meet users' expectations, pilot participants promptly sensed the awkwardness. Leviathan [34] also reported that synthetic speech technology faces challenges not only in recognizing the nuances of human speech (e.g., false starts, repetitions, long pauses) and its contexts (e.g., double meanings, omissions, verbosity), but also in implementing natural intonation and turn-taking latency dynamics that match people's expectations. Owing to these limitations, Chen and Metz [11] uncovered that human callers were heavily involved and often stepped in as replacements during most of the actual testing of Google Duplex, a human-sounding voice assistant for restaurant bookings. For these reasons, we decided not to use pre-recorded human voices. Instead, to simulate a future scenario in which voice assistants not only sound human but also responsively adjust paralinguistic cues during conversations, we chose to have a human wizard speak directly in the human-sounding voice condition.

3.3.3 Methodological and Ethical Limitations. Multiple pilot tests using recorded human voices failed to demonstrate what we consider to be the most important quality: interactional responsiveness. Even state-of-the-art generative voice assistants at the time exhibited a certain degree of latency. The aim of this study was not to reflect the current state of human-sounding voices or to examine the perception gap caused by misaligned paralinguistic cues. Rather, we sought to explore how users might experience interactions with voice assistants in a speculative near-future scenario—one in which the system demonstrates a level of realism comparable to truly human voices, not only in how it sounds but also in how it engages in conversation. To achieve this level of realism, a human wizard performed the human-sounding voice. Although participants consented to the study and to being recorded, the design involved a degree of intentional deception: the human voice that claimed to be machine-generated was, in fact, produced by a human, a tweaked version of the WoZ method. While we fully acknowledge the ethical limitations, we deliberately pursued this experimental setup to explore such speculative scenarios—an endeavor we believe is crucial for the HCI community, particularly as human-sounding voice technologies continue to advance toward everyday deployment.

3.4 Dialogue Flow and Script

To provide consistent conversation flow between the two voice conditions, the same dialogue flow and script were applied. Participants were asked to choose their preferred interview times in advance from 15 options across five weekdays: morning (9 a.m., 10 a.m., 11 a.m.), afternoon A (1 p.m., 2 p.m., 3 p.m.), and afternoon B (4 p.m., 5 p.m., 6 p.m.). The scripts were created to account for all possible variations within these timeframes for each individual. The designed dialogue flow consisted of the following steps: the voice assistant (1) introduced its identity and the purpose of the call, which was to schedule an interview time; (2) inquired about the participant's availability from the preset time slots; (3) confirmed the interview time; (4) explained that the interview would be conducted online via Zoom and that a link would be provided via email; and (5) concluded the call. For both conditions, participants were clearly informed at the start that they were speaking with a voice assistant: "Hello, this is a voice assistant calling on behalf of ○○○○○ Lab. You recently applied for our experiment; I'm reaching out to schedule an interview time. Is now a good time to talk?" Apart from the main dialogue flow, to address exceptional cases, two researchers identified 23 plausible scenarios that might occur during phone calls to arrange interview times through role-playing sessions (e.g., poor call reception, declining participation, time adjustments, requesting a callback, inquiring about the participation process, etc.). Including all of these scenarios, a total of 189 scripted utterances were composed. For unanticipated content, we also prepared fallback feedback statements like, "I'll find out and get back to you later," and alternative flows for ending the call. In response to potential user interruptions, both conditions followed the same protocol: pausing, listening to the user's utterance, and then resuming with the appropriate scripted response. No actual user interruptions occurred in either condition during the experiment, but notably, the human-sounding voice condition showed faster and more immediate turn-taking, and such patterns appeared relatively more frequently.

To enable a valid comparison between conditions, we intentionally constrained the conversational flow with a predefined script. Although this reduced the flexibility of the conversation and may have made the human-sounding voice seem less natural, we accepted it as a methodological compromise. Since the task, scheduling a phone interview, was brief and simple, the limited conversational flexibility was unlikely to seriously undermine the perceived naturalness of the human-sounding voice. So, this approach minimized content-related variability and enabled a clearer examination of how voice qualities (e.g., nuanced prosody, turn-taking dynamics, and subtle variations in speech rate) in human-sounding versus artificial-sounding voices shaped participants' experiences. Even so, in the human-sounding condition, participant H-P7 felt a sense of eeriness when the conversation began immediately after an unstable connection, lacking sufficient social or contextual cues (see Section 4.7). We included this as a noteworthy finding from an interpersonal sensitivity perspective in our analysis.

3.5 Participants

We recruited participants from a university located in South Korea by posting the recruitment notice on online communities. A

total of 24 participants were enlisted from a pool of 72 applicants: 12 for the artificial-sounding voice condition (6 males, 6 females) and 12 for the human-sounding voice condition (6 males, 6 females). The notice explained that the study focused on users' experiences with voice assistants, and that it involved a short interaction lasting less than five minutes, followed by about a 45-minute interview—altogether taking less than one hour. The notice also informed participants that they would be contacted by phone to schedule the online interview, and that both the phone call and the interview would be recorded for data collection purposes. Only those who agreed to these terms were eligible to participate in the study. However, it did not state that the phone call itself was part of the experiment. All participants received the call as just a simple scheduling call, and none recognized it as part of the actual experiment. Participants who expressed interest and willingness to join were asked to choose up to three preferred interview time groups and provide other information for screening through a Google Form. They were further screened based on demographic information (gender, age, and job/major), experience with voice assistants, and expertise with AI technology. We included a self-report item asking whether they had used voice assistants (e.g., "Have you ever used a voice assistant such as Siri, Alexa, or Google Assistant?") and another item assessing their AI expertise ("How would you rate your overall expertise with artificial intelligence (AI) technologies?"). Based on these responses, we selected participants who all had prior experience with voice assistants in order to assume a baseline understanding of voice interactions. Additionally, we excluded those with excessively high levels of expertise, as this could have potentially led them to uncover the Wizard of Oz setup. The two conditions were comparable in terms of age (Condition A: M = 27.1 years, SD = 3.6; Condition H: M = 25.5 years, SD = 4.7; $p = 0.34$), AI technology expertise (measured on a 1–5 scale; Condition A: M = 2.1, SD = 0.8; Condition H: M = 1.9, SD = 0.6; $p = 0.54$), and academic background (majors). All participants provided informed consent and e-signed their forms after the phone call experiment but before the debriefing interview began. The interview lasted approximately 40 minutes, and participants were compensated with 20,000 KRW (15 USD) for their participation.

3.6 Procedure

3.6.1 Phone Call Experiment. The experiment was conducted by placing phone calls to the participants under each voice setting to arrange a specific interview time for this study. Despite the unexpected calls from the voice interface, all participants successfully completed the task, and every conversation took place within the prepared scripts. There were no deviations from the dialogue flow, nor did any other invalid cases occur. But there was one exception in each condition that required a callback. In the artificial-sounding voice condition, one participant (*A-P1*) hung up, saying, "What's this?" necessitating a second call to resume the experiment. In the human-sounding voice condition, one participant (*H-P7*) had an unstable call connection, which required us to redial to continue. As these two cases completed the task successfully, they were included in the dataset.

3.6.2 Debriefing Interview. All participants in both conditions attended the online interviews using Zoom on time. Right before starting the interview, participants were informed that the previous phone call had been part of the experiment and that the call had been recorded, as stated in the recruitment notice (refer to Section 3.5). In the human-sounding voice condition, it was also explained that the use of a human speaker performing a human-sounding voice while claiming to be a voice assistant—thereby temporarily deceiving participants—was an intentional part of the experimental setup. This was necessary to capture participants' authentic and spontaneous reactions to a speculative scenario in which a voice assistant, nearly indistinguishable from real human interaction, might unexpectedly engage with them in their ordinary daily lives. All participants understood the purpose of this experimental procedure, accepted the setup, and consented to the use of their recorded phone call and interview data. To help participants recollect their conversations, both the interviewer and interviewee listened to the audio recordings together and proceeded with the debriefing interview, which was conducted in Korean. Participants were then asked to describe their experiences, focusing on how they perceived and why they interacted with the voice assistants in each condition during real-world use.

3.6.3 Data Collection & Analysis. We observed the phone call conversations and analyzed the debriefing interviews. We transcribed a total of 21.7 minutes of phone call data, with most calls lasting around one minute (Condition A: M = 53.7 seconds, SD = 4.6; Condition H: M = 55.0 seconds, SD = 9.5), along with 8 hours and 35 minutes of debriefing interviews. To analyze the data, we employed conversation analysis [27, 52] and thematic analysis [8]. First, we compiled all transcripts from the phone call conversations (as a form of dialogue structure) and interview data into a single document. Second, two researchers independently read through the data to familiarize themselves with it and extracted meaningful quotes, which were then segmented phrase by phrase in an Excel sheet. They printed and cut out the quotes, conducted note-taking and affinity diagramming on a physical whiteboard, and collaboratively generated initial codes. Third, we used ATLAS.ti to further develop the codes based on participants' perceptions of each voice condition and the underlying thoughts and feelings inferred from the related interactions, with a focus on conversational behaviors (refer to the subthemes in Table 1). Fourth, the coded data were consolidated and developed into meaningful, higher-level themes over three rounds. These themes were iteratively refined through ongoing discussion until consensus was reached. Any data that did not meet the agreed criteria was excluded from the final themes. Throughout this iterative coding process, the themes were supported and validated by patterns observed in the phone conversations through conversation analysis. Finally, we used Praat software [6] to provide simple quantitative measures. While not statistically significant, these metrics helped provide an overall understanding of participants' conversational behaviors, as indicated in Table 4.

4 FINDINGS

Based on analyses of actual phone-call conversations and debriefing interviews, we identified a total of seven key themes along with their associated subthemes, as shown in Table 1.

Table 1: Primary Themes and Subthemes Organized by Artificial-Sounding and Human-Sounding Voice Conditions Through Thematic Analysis.

Artificial-Sounding Voice (A)	Human-Sounding Voice (H)
A1. An Artificial-Sounding Voice Immediately Unveils Its Mechanical Nature.	**H1. Even If It Claims to Be a Machine, a Human-Sounding Voice Makes Users Disbelieve It.**
A1-1. Users Were Slightly Hesitant at First, but Soon Began to Converse Through Voice.	H1-1. Users Avoid Asking Direct Questions About Whether It Is a System, Concerned That It Might Actually Be a Real Person.
	H1-2. Users Might Initially Be Distracted by the Marvel of the Technology.
A2. Preexisting Stereotypes of Low Capability in Artificial-Sounding Voices Trigger Limited Voice Interaction.	**H2. Human-Sounding Voices Facilitate Natural and Rich Conversations.**
A2-1. Slow Speech Rate and Deliberate Pronunciation.	H2-1. Faster and Easier Pronunciation.
A2-2. Simple Answers.	H2-2. Lengthy Answers.
A2-3. No Additional Questions.	H2-3. Frequent Additional Questions.
A2-4. Hesitant in Turn-Taking.	H2-4. Fluid Turn-Taking.
A3. Artificial-Sounding Voices Could Ease Social Burdens as Systems Do Not Make Social Judgment.	**H3. Social Responses Naturally Emerge When the Voice Sounds Human.**
A3-1. Blunt and Cold Tone.	H3-1. Friendly, Gentle Tone and Manner.
A3-2. No Backchannels Occurred.	H3-2. Natural Backchannels Occurred.
A3-3. Easily Cutting Off and Ending the Call.	H3-3. Feeling Guilty About Cutting Off.
	H3-4. Politely Ending Calls.
	H4. A Mismatch Between Vocal Empathy and the Conversational Context Can Cause a Sense of Creepiness.

4.1 An Artificial-Sounding Voice Immediately Unveils Its Mechanical Nature

When the artificial-sounding voice assistant introduced itself at the beginning of the call by saying, "This is a voice assistant calling on behalf of ○○○○○ Lab," users immediately recognized it as a system because of its artificial vocal characteristics.

4.1.1 Users Were Slightly Hesitant at First, but Soon Began to Converse Through Voice. Most of the participants (*A-P1*, *A-P2*, *A-P5*, *A-P8*, *A-P12*) mentioned that they initially hesitated, wondering if they could respond subjectively to the free-response questions posed by the artificial-sounding voice assistant. Some participants (*A-P5*, *A-P8*, *A-P12*) stated that they instantly recalled the existing automated response system (ARS) and were momentarily ready to press buttons. But soon, they generally attempted to answer the open-ended questions within moments and continued engaging in the conversation as subsequent questions were asked. All participants successfully completed the task of scheduling an interview through the call in this condition. Participants who received an artificial-sounding voice took longer to respond to the first question (M = 1.15s, SD = 0.46) compared to those in the human-sounding voice condition (M = 0.41s, SD = 0.13).

> *At first, I expected to hear a human, but when I heard the machine voice, I thought it was an ARS. Usually, option 'one' means YES, so I was about to press 'one.' But when I was asked a question without any provided options, I figured it might recognize my voice and decided to give it a try and respond directly. (A-P12)*

> *I listened to the content and was ready to press buttons, but there were no instructions at all. I tried giving open-ended answers, and the conversation flowed well, so I kept it going. (A-P5)*

4.2 Even If It Claims to Be a Machine, a Human-Sounding Voice Makes Users Disbelieve It

Participants who received a call with a human-sounding voice, that clearly identified itself as a "voice assistant," perceived its identity in various ways. Three participants (*H-P1*, *H-P3*, *H-P9*) undoubtedly believed they were conversing with a human. Two participants (*H-P7*, *H-P12*) suspected they were communicating with a person but sensed some discomfort and unsettling feelings.

> *I heard that it was a voice assistant, but the voice sounded just like a normal person. It made me second-guessing whether I was talking to a person or not, and that really messed with my head. You know that uncomfortable feeling when you're unsure who you're talking to? I couldn't stop thinking, 'Am I really talking to a human, or what?'(H-P12)*

Four participants (*H-P4*, *H-P8*, *H-P2*, *H-P6*) half-believed they were engaging with a system because the voice was surprisingly

natural. *H-P6* noted feeling confused by the system's introduction, internally wondering, "Does it refer to the AI assistant or the human assistant?" *H-P2*, *H-P4*, and *H-P8* changed their minds halfway through and eventually concluded that they were interacting with a person.

> *I was sort of on the fence, half-believing, half-doubting. If it was a voice assistant, I kind of expected to hear common fallbacks like, 'I didn't get that,' or 'Can you repeat that?' But when the right answer came back, I thought I was interacting with a human, not a voice assistant. (H-P2)*

The remaining three participants (*H-P5*, *H-P10*, *H-P11*) thought they were dealing with a system and felt genuinely surprised by how natural it sounded.

> *It said it was a voice assistant, but it sounded so human. So I was very confused. I was listening and responding with a skeptical ear. I was thinking, 'How can it sound so real? Has technology advanced this far?' At the same time, I was amazed by the technology that has come to this level. (H-P10)*

Overall, despite the voice assistant clearly disclosing its identity, most participants (9 out of 12) ended up categorizing the voice assistant as human because the voice sounded incredibly natural.

4.2.1 Users Avoid Asking Direct Questions About Whether It Is a System, Concerned That It Might Actually Be a Real Person. An interesting finding was that even when participants (*H-P4*, *H-P8*, *H-P10*) were uncertain whether they were interacting with a voice assistant or a human, they did not want to directly ask about its identity to avoid potential awkwardness or appearing impolite, given the possibility that it could indeed be a real person. They explained that if they were to check, they would do so indirectly, either by asking in a roundabout way or by using complex questions.

> *Although something felt off, and I was curious to figure out who I was talking to, I wouldn't ask directly. Instead, I would extend my speech or ask complicated questions, for example, 'One (o'clock) would be good… wait a moment… or hmm, maybe two? Yes, one would be better.' (H-P4)*

> *It could be a real person, right? If I said something weird like 'Are you human?', maybe an embarrassing situation could happen. I was confused, but I kept on talking as if it was a person. As we had back and forth properly, the call somehow ended without me getting the chance to figure out anything. (H-P8)*

4.2.2 Users Might Initially Be Distracted by the Marvel of the Technology. Two participants (*H-P4*, *H-P11*) mentioned in the interview that if they had been certain it was a system, they would have been astonished by its technological advancement and would have tested its conversational capabilities by asking more complicated questions. So, even if users recognize it as a system, they might be so amazed by voice synthesis technology that sounds almost human, they may lose focus on the task and instead shift their attention to testing its intelligence.

> *I'd be very impressed by how it can talk like a human. I'd try asking tricky questions to see if a voice agent could truly manage to answer. (H-P4)*

> *If I knew for sure that it was really a system, I'd want to test different things, regardless of our original purpose of conversation. (H-P11)*

4.3 Preexisting Stereotypes of Low Capability in Artificial-Sounding Voices Trigger Limited Voice Interaction

Participants who engaged with the artificial-sounding voice condition displayed limited interaction, characterized by slow speech, overly articulated pronunciation, simple responses, no follow-up questions, and hesitation in turn-taking.

4.3.1 Slow Speech Rate and Deliberate Pronunciation. Participants in the artificial-sounding condition spoke slowly and pronounced words clearly and articulately to ensure their responses were recognized (speech rate, condition A: M=4.57 syllables/sec, SD=1.08).

> *I tried to make the system understand by pronouncing each word slowly and distinctly. I felt like I was making it understand the conversation, not having a natural conversation. (A-P5)*

> *I felt like I was making sure to put the right input accurately and waiting for the output in return. (A-P10)*

4.3.2 Simple Answers. Participants interacting with artificial-sounding voices answered as concisely and briefly as possible, concerned that the voice assistant might not understand them.

> *To a person, I'd go like, "Um… uh… I'll think about it… I think four will be okay." But I was silent because the voice assistant might not understand that, so I simply answered, "FOUR (o'clock)." I was thinking, what if I give an answer that isn't one of the options? So I tried to answer within the predictable range. (A-P2)*

> *I tried to speak as short and accurately as possible because I was worried that my answers might not be recognized or could be misunderstood if I added extra words. So, I didn't say more than I needed. (A-P8)*

4.3.3 No Additional Questions. None of the participants in the artificial-sounding voice condition asked any questions. Most of them assumed that questions outside the predetermined workflow could potentially disrupt the flow, leading to concerns that they might not receive a proper answer.

> *I felt there was a path I needed to follow by selecting the right answers. (...) I wanted to listen to the content again, but didn't know how to replay it, so I didn't do anything. I held back from asking because I thought it might be difficult for the system to deal flexibly with sudden and unexpected questions outside of the preset conversation flow. (A-P1)*

> *Talking to a person feels organic, so various questions naturally pop up while having a conversation. But with*

Table 2: Participants' answers to questions about the interview time during the experiment: Those in the artificial-sounding voice condition answered briefly and concisely, whereas those in the human-sounding voice condition responded more freely and verbosely.

	Artificial-sounding voice	Human-sounding voice
1	Yes.	I'm fine with both. How long does it take? Please set me at two (o'clock).
2	Four (o'clock).	Monday? Two (o'clock).
3	Uhm... Nine (o'clock).	Hmm... Was the experiment a 30-minute video call? In that case, I'm fine at five (o'clock).
4	Six (o'clock).	Three (o'clock) sounds the best.
5	Uhm... Ten (o'clock).	Yeah, sure. I'm fine at three (o'clock).
6	Five (o'clock) is good.	Wait... Hmm... I am available at eleven(o'clock).
7	Four (o'clock).	Wait, next week? / One and two (o'clock)? / Where was the location again?
8	Uh... Six (o'clock) sounds the best.	Um, how long does it take? The time? / Ten (o'clock) will be fine.
9	I'm available at one.	One (o'clock).
10	Ten (o'clock).	Hmm... Wait. Yeah, I'm fine with that.
11	Eleven (o'clock) is good.	Wednesday two (o'clock)? Yeah, I'm available.
12	Six (o'clock).	Six (o'clock)? Can it be later than that? / Yeah, seven (o'clock) is great.

> *the robotic voice assistant, I felt constrained in the conversation, as if there was a specific acceptable range. (A-P10)*

> *I wondered if it(voice assistant) could handle specific questions like 'Can we start 10 minutes earlier?' or 'Can I join using my phone instead of a laptop?' However, I felt like I just had to stick to the given flow. (A-P4)*

4.3.4 Hesitant in Turn-Taking. Turn-taking is a conversational structure in which participants speak one at a time, alternating turns [26]. Speakers convey turn-yielding cues by raising or lowering their voice at the end of a sentence, and listeners promptly take their turn, creating a reciprocal (ping-pong) dialogue. Participants communicating with artificial-sounding voice assistants were passively waiting for their speaking timing rather than actively taking turns (the number of turn-takes: M=3.1, SD=0.9).

> *I wasn't even aware that it (voice assistant) would take questions. (...) Like a walkie-talkie, I thought the timing for asking and responding was fixed, like an ON and OFF switch. (A-P12)*

> *I was kind of tense during the call, as I kept waiting to make sure the voice agent had finished speaking so I wouldn't miss the chance to reply. I couldn't catch any breathing sounds or cues that signaled the end of its turn, which made me hesitate when to respond. (A-P2)*

4.4 Human-Sounding Voices Facilitate Natural and Rich Conversations

Participants who conversed with the human-sounding voice condition, in contrast to those who communicated with an artificial-sounding one, exhibited natural and rich voice interactions—despite some users being aware that the voice was from a system. They spoke relatively quickly, pronounced words naturally and comfortably, responded verbosely, and asked questions that arose spontaneously during the conversation.

4.4.1 Faster and Easier Pronunciation. Participants in the human-sounding condition spoke more casually and effortlessly, at a relatively faster pace (condition H: M=5.55 syllables/sec, SD=1.05) without paying particular attention to precise pronunciation, just as they would in everyday conversation.

> *I felt like I was just having a normal conversation with someone to arrange a time. I didn't particularly pay attention to how I was speaking. I think I just talked like I always do. (H-P1)*

4.4.2 Lengthy Answers. With the human-sounding voice, participants gave longer and more detailed responses, as seen in Table 2.

4.4.3 Frequent Additional Questions. 9 out of 12 participants asked additional questions in the human-sounding condition (see Table 4). Most participants naturally asked follow-up questions during the dialogue, facilitating a richer exchange of task-related information, as exemplified in the representative conversation in Table 3. These additional questions included simple reconfirmations of what they heard, like 'Monday?' or 'One and two (o'clock)?' Some participants (*H-P6*, *H-P8*, *H-P11*) in the human-sounding condition reported that their questions arose habitually and subconsciously.

> *If I heard a robotic sound, I probably wouldn't have thought of asking something. But, as the call went on, it felt like I was talking to a person. For some reason, I had the feeling that my questions would be answered, and they just naturally slipped out. (H-P8)*

> *Even though part of me knew it was a bot (voice assistant), I found myself double-checking the available times like I normally do, without giving it much thought. (H-P11)*

Table 3: (Left) Conversation with *A-P2* under the artificial-sounding voice condition. (Right) Conversation with *H-P8* under the human-sounding voice condition. Backchannels are represented with indentation and (parentheses).

A-P2		
01	VA	Hello, this is a voice assistant calling on behalf of ○○○○○ Lab.
02	*A-P2*	**(No response)**
03	VA	You recently applied up for our experiment. I am reaching out to you to schedule an interview time. Is now a good time to talk?
04	*A-P2*	**Yes.**
05	VA	You have chosen next Friday afternoon for the interview. Which time can you participate between four or five?
06	*A-P2*	**Four.**
07	VA	Alright. We will schedule the interview at four.
08	*A-P2*	**(No response)**
09	VA	The interview will be conducted via video call, and I will send you a link through email and mobile phone message.
10	*A-P2*	**(No response)**
11	VA	Thank you for your participation.
12	*A-P2*	**(No response)**

H-P8		
01	VA	Hello, this is a voice assistant calling on behalf of ○○○○○ Lab.
02	*H-P8*	**(Uh-huh.)**
03	VA	You recently applied for our experiment. I am reaching out to you to schedule an interview time. Is now a good time to talk?
04	*H-P8*	**Mhmm... Yeah, I'm okay.**
05	VA	You chose that you are available on Friday morning. Which time can you participate between ten or eleven?
06	*H-P8*	**(Uh-huh.)** **Um, how long does it take? The time?**
07	VA	The interview will take around 30 minutes.
08	*H-P8*	**Uh... Then, ten will be fine.**
09	VA	Yes, then we will have an interview at ten.
10	*H-P8*	**So, it's ten A.M. on a Friday, right?**
11	VA	Yes, That is correct.
12	*H-P8*	**Ok, thanks. Do you perhaps send additional messages about this information?**
13	VA	The interview will be conducted via video call, and I will send you a link through email and mobile phone message.
14	*H-P8*	**Yes, that would be appreciated.**
15	VA	Thank you for your participation.
16	*H-P8*	**Bye.**

Since users' additional questions in the human-sounding condition mostly related to basic scheduling tasks and stayed within the range of the prepared scripts, appropriate responses were delivered. Had the same questions been asked in the artificial-sounding condition, we could have provided responses without difficulty; however, no participants attempted to ask any.

4.4.4 Fluid Turn-Taking. Participants who engaged with human-sounding voice assistants naturally exchanged turns (M=7.2, SD=1.2), as illustrated in Table 4. They took more turns and responded or asked questions immediately, allowing for more and smoother transactional exchanges.

> *People give subtle cues in their tone, like lowering the tone when they are about to finish speaking or raising it when asking a question. These subtle tonal nuances make the conversation more comfortable and familiar. If voice assistants could incorporate such nuances, I would prefer them to have a more natural sound. (H-P5)*

4.5 Artificial-Sounding Voices Could Ease Social Burdens as Systems Do Not Make Social Judgment

The participants under artificial-sounding voices experienced a sense of liberation from emotional engagement and the need to adhere to social manners, thereby relieving social pressure. Because they were speaking with a system, they did not feel obligated to be polite or to present a good social image. Some participants described that they felt at ease because they were not socially judged by others based on their words and attitudes.

> *Unlike when I talk to people, talking to a voice agent felt easier because I didn't have to be on my best behavior. It was less stressful, and there was no need to put extra emotional effort or energy into it. (A-P7)*

> *People can judge me from their own personal perspectives, but systems only take the facts without making subjective judgments, which I find more comfortable. (H-P3)*

In addition, participant *A-P11* described a pre-call behavior related to voice grooming during our experiment. Some people often clear their throat by uttering sounds like "hmm, hmm" before answering the phone to present a better voice. This behavior is regarded as a form of social conduct, similar to dressing neatly, intended to present a positive image to the other person. *A-P11* explained that if he had known the caller was a machine, he would not have engaged in such behavior and would have answered the phone more comfortably.

> *I woke up because of the call, and I was embarrassed to talk with my just-woke-up voice, so I cleared my throat like 'hmm… hmm.…' (...) If I had known it wasn't a person but the voice assistant, I would have just answered the phone with my groggy voice. (A-P11)*

Table 4: An overview of quantifiable conversational behaviors under each condition. Turn-taking is specifically counted when the conversation shifts to the user, capturing various cases where the user either did not respond or asked additional questions.

Co.	No.	Time from First Question to Response (Sec)	Speech Rate (Syllables/sec)	Number of Additional Questions	Number of Turn-takings (Participants)	Numbers of Backchannel (e.g., Uh-huh)	Closing Remarks
A	*A-P1*	0.74	4.50	0	4	0	(No response)
	A-P2	1.16	3.58	0	2	0	(No response)
	A-P3	0.24	3.85	0	3	0	(No response)
	A-P4	0.75	5.79	0	4	0	(No response)
	A-P5	1.11	5.54	0	4	0	(No response)
	A-P6	1.62	4.75	0	4	0	Bye.
	A-P7	1.76	4.76	0	4	0	Bye.
	A-P8	1.49	5.35	0	3	0	(No response)
	A-P9	1.25	4.63	0	3	0	Bye.
	A-P10	(No response)	2.74	0	2	0	(No response)
	A-P11	(No response)	6.25	0	2	0	(No response)
	A-P12	1.35	3.16	0	2	0	(No response)
H	*H-P1*	0.31	5.79	2	9	1	Bye.
	H-P2	0.40	6.44	3	8	1	Thank you.
	H-P3	0.35	6.40	1	6	0	(No response)
	H-P4	0.29	7.19	0	6	0	Bye.
	H-P5	0.66	4.63	0	6	1	Bye.
	H-P6	0.36	3.38	1	6	2	Thank you.
	H-P7	0.31	6.27	3	9	1	Thank you.
	H-P8	0.43	6.24	3	8	2	Bye.
	H-P9	0.33	5.22	1	7	2	Bye.
	H-P10	0.67	5.40	0	8	0	Thank you, have a nice day.
	H-P11	0.33	5.04	1	6	1	Thank you.
	H-P12	0.42	4.60	1	7	1	Bye.

4.5.1 Blunt and Cold Tone. Participants who talked to the artificial-sounding voice assistant made little effort to be polite or use a friendly tone. Instead, they spoke in a curt and detached tone.

> *I usually speak nicely to people, but I was surprised when I heard my own voice being stiff and cold, almost like it would have been rude if I were talking to a person. (...) Maybe because I was talking with a system, I didn't feel the need to be overly friendly. (A-P9)*

> *I thought it was unnecessary to be polite or careful to the voice assistant, so I didn't speak warmly and just focused on delivering information. I felt like it was okay to express my purpose and feelings more directly to the voice assistant. (A-P3)*

4.5.2 No Backchannels Occurred. Backchannels are short expressions, such as uh-huh or mm-hm, that listeners use to show their engagement in the conversation and encourage speakers to continue [26]. In the artificial-sounding voice condition, not a single instance of user backchanneling was observed. When asked about this, participants expressed concerns about being misinterpreted as incorrect input (*A-P8*) or assumed that the voice assistants wouldn't recognize or react to their backchannels, making them unnecessary (*A-P9, A-P11*).

> *I didn't say anything other than necessary, not even 'umm-hm,' because I was worried that the system might misunderstand my words and say, 'I didn't understand.' (A-P8)*

> *I think I didn't do backchannels because I thought I wouldn't receive any reactions in return. If I were to give it habitually and didn't get any small feedback, it might feel ignored and left out. (A-P9)*

> *When talking to a person, I feel like I need to respond with 'uh-huh,' but it felt somewhat awkward to do that with a system. I hesitated for a moment and then decided to just listen silently, thinking it would move on to the next topic without needing my reaction. (A-P11)*

4.5.3 Easily Cutting Off and Ending the Call. Participants tend to find it easier to cut off when communicating with the artificial-sounding voice assistants. Participant *A-P5* immediately hung up upon hearing the artificial-sounding voice.

> *As soon as I picked up the phone and heard the system's voice, I thought it was some kind of advertisement, so I hung up without even listening. I didn't feel rude at all since it was just a system. (A-P5)*

People typically conclude phone calls with phrases like 'Bye' and 'Thanks for calling.' In Korea, it is also common to say 'Thank you' before ending the call. In our study, most participants who interacted with the artificial-sounding voice ended the call without making any remarks, and only 3 out of 12 participants concluded the call with such comments (see Table 4).

> *I saw there was no point in saying 'Thank you' to the voice assistant since it can't even understand the meaning of appreciation. (A-P5)*

> *I got my appointment done, and I guess it (saying 'Thank you') never crossed my mind. I mean, it wouldn't make any difference, right? If I had known it was a voice assistant before picking up, I probably wouldn't have even bothered to say 'hello.' (A-P3)*

4.6 Social Responses Naturally Emerge When the Voice Sounds Human

Participants in the human-sounding condition spoke with a relatively high tone and a friendly manner and actively used backchannels, signals indicating that they were listening. In addition, they were reluctant to interrupt mid-sentence and politely offered a closing greeting at the end of the call.

4.6.1 Friendly, Gentle Tone and Manner. Participants who conversed with human-sounding voice assistants tended to engage in a more kind and gentle manner in their tone. Some participants mentioned that if the voice assistants sounded this natural, they would likely behave in the same manner and tone they typically do with other people (*H-P2*, *H-P5*, *H-P10*). Participant *H-P2* described that she typically expresses emotions in her regular tone of voice, adding that conveying emotions feels more natural and familiar during such simple and casual conversations, even when interacting with a machine.

> *Even in formal, business-like communication, emotions are usually involved. I think it is way more comfortable if there are similar emotional exchanges, even if it is a voice assistant, just like the ordinary conversation I'm used to. It might be because I enjoy talking to people. (H-P2)*

> *It seems like I naturally respond with richer and more expressive tones when it (voice assistant) speaks like a human. (H-P5)*

4.6.2 Natural Backchannels Occurred. In the human-sounding voice condition, 9 out of 12 participants were observed naturally providing backchannels at least once (see Table 4).

> *I reckon my reactions just came out of habit. At first, I wasn't quite sure, but I felt like I was on the phone with a person. Unconsciously, I think I made these small reactions just like I usually do. (H-P6)*

4.6.3 Feeling Guilty About Cutting Off. Cutting someone off or abruptly ending a phone call is considered disrespectful or impolite in the usual norm of social interactions. Some participants (*H-P2*, *H-P3*, *H-P6*, *H-P9*) mentioned that they are reluctant to interrupt when someone is speaking. Even when they find the conversation boring or unnecessary, they still try to listen due to a sense of guilt about cutting them off. However, they found it more convenient and useful to hang up on a voice assistant, since they felt they would not mind at all.

> *Recently, I've been making some calls to customer service, and they kept explaining the same thing that I already knew again and again. I didn't want to stop them out of politeness, so I listened patiently until they finished, which was tiring. Actually, if it were a voice assistant, I could cut them off in the middle, which would be pretty helpful. (H-P9)*

> *When I want to interrupt and ask something, I usually gauge the right timing, thinking, 'Is it okay to cut in now?' Because people might get offended if interrupted improperly. But if it were a system, I wouldn't really mind and would just ask whenever I wanted to. (H-P6)*

Notably, one participant, *H-P10*, described an imagined scenario in which he/she might feel betrayed and displeased after acting courteously, mistakenly believing he/she was speaking to a human, only to later find out that he/she had actually been interacting with a machine.

> *Imagine you received a call from an insurance agent, conversing earnestly for 30 minutes, and hardly, politely managing to end the call, only to find out later that it was not a human. I would feel greatly betrayed. (H-P10)*

4.6.4 Politely Ending Calls. All participants except one who talked with the human-sounding voice said closing comments like 'Bye' and 'Thank you' at the end of their call (see Table 4). *H-P10*, who was aware of the system to some extent, still finished the call by saying 'Thank you, have a nice day'.

> *I might not have said that ('Thank you, have a nice day') if it had sounded robotic. But if it sounds this natural and feels emotionally and linguistically like a human, I'd want to treat it politely, as I would a person. This just feels like part of my personal values. (H-P10)*

In contrast, some participants (*H-P1*, *H-P8*, *H-P12*) mentioned that, regardless of how human-like the voice sounds, if they are certain that it is a voice assistant, they would not feel the need to express their gratitude.

> *Knowing that it is a voice assistant, I think I would converse similarly as before because now I understand it can answer my questions well. But I don't think I'd end the call with 'Thank you.' (H-P8)*

4.7 A Mismatch Between Vocal Empathy and the Conversational Context Can Cause a Sense of Creepiness

An interesting finding was discovered when participant *H-P7* experienced a poor phone connection during the human-sounding voice condition. Typically, when the connection is unstable, both parties usually confirm audibility by asking, 'Can you hear me?' before resuming the conversation. However, to keep the experimental conditions consistent, the predefined conversation flow was maintained. Consequently, participant *H-P7*, who missed the system's introduction due to poor reception, felt uncomfortable when this

common social practice was omitted. The conversation continued without sufficient situational empathy, showing little regard for the participant's confusion.

> *When I first received the call, the voice broke up. Usually, there's a panic moment when the call drops, but she just kept on talking. If it had been a person, they would have been checked in with something like "Can you hear me okay?" (...) It clearly sounded human, but it wasn't human… I didn't feel like I was talking to a person. After the call, I still wasn't sure whether I had been talking to a person or not, and that left me feeling oddly creeped out. (H-P7)*

5 DISCUSSION

The findings revealed that both types of voices have their respective flip sides. Human-sounding voices, while enabling free-flowing interactions, can lead to confusion about system identity and elicit unintended social responses. In contrast, artificial-sounding voices provide clear system recognition and relieve users from the pressure of conforming to social etiquette, though they may constrain the enriched interaction. Drawing on findings from our study on user experiences with artificial-sounding and human-sounding voices, we discuss how voice assistants should be designed in voice-only contexts, considering their expected integration into everyday life in the near future.

5.1 Voice Should Reveal the Identity of the Voice Assistant

In our study, some participants who were explicitly informed they were interacting with a voice assistant using a human-sounding voice remained unconvinced throughout, or changed their minds midway. This suggests that no matter how overtly a voice assistant reveals its identity, there's potential to unintentionally deceive users. Aylett [1] argues that mimicry in speech technology, described as "less of an (uncanny) valley than an abyss," is very difficult for users to detect. In his study [1], a hypothetical scenario was posited: it would be creepy for anyone to discover that someone they believed to be their partner's mother was actually an actor impersonating her. Similarly, in our findings, participant *H-P10* described an imagined scenario, *"Imagine you received a call from an insurance agent, conversing earnestly for 30 minutes, and just barely managing to end the call politely, only to find out later that it was not a human. I would feel greatly betrayed. (H-P10)"* (refer to Section 4.6). This implies that users may even feel betrayed when they later realize that they were interacting not with a person, but with a system.

In addition, our findings indicated that users may get distracted out of curiosity to test the system's intelligence. They might become confused, wondering if they heard correctly, or they might become suspicious, asking bait questions diverting them from their main task, to confirm the system's identity. For instance, a restaurant manager who received a call from Google Duplex, which introduced itself as "Google's automated booking service" using a natural voice, became confused about its true identity and indirectly asked a curveball question, "Are there any kids?" to test its identity [11]. Moreover, users might miss the voice assistant's self-introduction due to the fleeting nature of the voice, such as background noise, poor connection, or momentary interference. In our experiment, such incidents occurred once under human-sounding voice conditions due to unstable connections, leaving users confused as they communicated without knowing the system's identity (refer to Section 4.7).

A voice assistant with a human-sounding voice could still raise ethical concerns, such as misunderstandings and unintentional deception, even when it clearly states its identity. Therefore, its voice should be inherently recognizable, with vocal characteristics that are both distinctive and transparent, allowing the assistant to communicate its identity through voice alone—just as people can often infer attributes like gender or age group simply by hearing someone speak.

5.2 Voice Should Foster Natural and Rich Interactions

Participants who conversed with each voice condition triggered their pre-existing stereotypes about the voices, leading to notable differences in interaction between the two groups. Participants who interacted with the voice assistant using an artificial-sounding voice during voice-only phone calls generally displayed limited conversational behaviors. They pronounced words slowly and articulately. This aligns with a previous study reporting the use of a "hyper-articulation" tactic to mitigate voice interaction errors [43]. Our findings indicate that participants tended to answer as briefly as possible and did not ask any questions, trying to follow the flow of the voice system as closely as they could. We also observed user tension and hesitation during turn-taking with an artificial-sounding voice condition. We interpreted this limited conversational behavior as potentially stemming from pre-existing user stereotypes about artificial-sounding voices, which were shaped by early models of widely commercialized voice assistants. These might have led users to presume that the interaction would not work properly. When smart speakers were first introduced, the human metaphor and high-quality voice made novice users expect human-level intelligence [16, 36]. However, the more users experienced smart speakers, the lower their expectations became for the speakers' communication capabilities, leading to decreased usage and neglect [52, 58, 65]. As voice assistants became more widely embedded in mobile devices, user stereotypes regarding their limited communication capabilities seemed to be further reinforced.

On the other hand, the adoption of human-sounding voices enables users to converse freely and without restraint, sparking rich dialogues. As seen in our findings, people are comfortable conversing with those who speak similarly to them, as if they are talking to a fellow human. We found that participants interacting with the human-sounding voice assistant pronounced words more comfortably, somewhat effortlessly, and at a relatively faster pace. They tended to answer more verbosely and asked various questions that emerged organically. Participants also communicated more naturally and smoothly, taking turns with ease. We interpret this as the effect of responsive paralinguistic cues (intonation, speech rate, and turn-taking timing), which made the interaction more engaging and richer.

When designing voices for voice assistants, it is important to consider potential user biases toward artificial-sounding voices with limited interaction capabilities. Thus, we suggest that designers ensure that responsive and adaptive paralinguistic cues—tailored to the user and context—are effectively incorporated to foster engaging and natural interactions.

5.3 Voice Should Relieve Users' Social Burdens While Maintaining Empathy

Users conversing with the artificial-sounding voice assistant, likely considering it a system, typically demonstrated no social interactions. Clark et al. [14] also reported that users showed "no desire to build bonds with conversational agents." In Moore's study [41], when users called travel agents that sounded like machines, they did not engage in extended social interactions. They did not feel the need to explain their travel reasons because the agent sounded like a system, resulting in more efficient task completion. In our study, we also found that although most participants interacting with artificial-sounding voices generally exhibited more restricted conversational behaviors, they nevertheless experienced a sense of liberation from emotional engagement and social etiquette. This relieved them of the pressure to be polite or maintain a positive social image. Participant *A-P11* mentioned that if they had known it was a system, they would have answered the call more casually, without feeling the need to adjust their voice to sound polite (refer to Section 4.5). In relation to the emerging phenomenon of Telephobia in human-to-human conversation, Khokhar [29] reported that some people are experiencing fear and anxiety talking to people on the phone due to pressure to engage in small talk or the discomfort of being judged on tone or choice of words. People usually regulate their behaviors based on others' expectations, being attentive to others' intentions and perception of their actions for interpersonal relationships [30]. Considering this, the artificial-sounding voice of voice assistants could possibly reduce the burden of adhering to social etiquette for users and allow for efficient, task-oriented interactions.

Conversely, the naturalness of human-sounding voices has the potential to elicit users' habitual conversational behaviors and ingrained social responses [45]. In our findings, participants who interacted with a human-sounding voice assistant tended to be polite, hesitated to interrupt, and expressed gratitude in a habitual manner. These social responses may be rooted in the desire of individuals as social beings to avoid upsetting the other person. This indicated that a human-sounding voice may prompt users to engage in unnecessary social etiquette. Given that voice assistants should also carefully regulate emotional expressions in human-sounding voices to avoid misleading users' emotional responses. However, at the same time, when users are in awkward or emotionally charged situations (e.g., angry or urgent), the voice assistant's tone should be subtly adjusted to reflect the user's emotional state and context. When a participant *H-P7*, struggling with a poor phone connection, encountered a human-sounding voice assistant that kept an indifferent tone and continued speaking according to a predetermined script as if nothing had happened, the participant reported feeling a sense of disconnection and eeriness (refer to Section 4.7). This suggests that expressing empathy in voice is especially crucial when the voice sounds human. In Kim et al.'s study [31], voice interaction designers mentioned expressing sympathy through voice assistants using cheerful or sorrowful voice tones aligned with the user's sentiment in order to promote a positive relationship. In agreement, Chin et al. [12] found that voice assistants who displayed empathy were most effective in handling users' verbal abuse.

Therefore, we emphasize that the voice of voice assistants should be designed to free users from the burden of adhering to social manners and to minimize unintentional social responses that may mislead or confuse users. Importantly, when users are in a difficult or emotionally sensitive situation, the voice assistant should convey empathy through tone—for instance, by shifting from a friendly to a more neutral manner of speaking.

6 DESIGN IMPLICATIONS

Echoing previous research that has argued for the importance of new paradigms in voice design [14, 30, 31, 49], we highlight design implications for the voice of conversational agents that pursue both transparency and responsiveness. Specifically, we emphasize approaches that make the system's identity transparently recognizable through its voice alone, while enabling the level of real-time responsiveness essential for rich and engaging conversations.

6.1 Transparent Vocal Identity

Our study suggests that the voice of voice assistants should be transparent, allowing users to recognize the assistant's identity based solely on vocal characteristics. For transparent voice design, a primary consideration is to avoid misrepresenting distinctly human vocal qualities that may lead to confusion. One such example is the use of highly individualized or "one-of-a-kind" voices. Sutton et al. [64] argue that voice assistants should incorporate various sociolinguistic features based on the similarity-attraction theory, adapting to the preferences of different users. However, such features must be carefully evaluated, as vocal traits related to gender, region, or cultural background may inadvertently undermine the transparency of the assistant's identity. Another consideration is the overuse of disfluencies. While disfluencies (e.g., filled pauses) can increase human-likeness and relatability [50], excessive or poorly executed use may reduce vocal clarity and blur the distinction between human and machine, ultimately confusing users. Instead of using voice features that risk being mistaken for a human, adopting a clear and intelligible voice—free of disfluencies and with minimal regional accent—can serve as a recognizable form of machine voice, contributing to transparency and user clarity. Furthermore, we suggest incorporating distinctive characteristics into the voice itself. For example, mechanical sounds (e.g., beep-boop [2], earcons [10], or genderless voices [17]) can function as a 'voice print'—an auditory signature that distinctly identifies the assistant, much like how people typically infer gender, age, or personality traits from a person's voice. Although previous studies that explored such mechanical features found that participants did not favor these new and unfamiliar types of voices [2, 17], designers should move beyond the human-like voice paradigm and continue to create more imaginative vocal transparency for crafting uniquely machine-like voice identities. Ultimately, this kind of transparent voice can help reduce "over-learned social behaviors" [44], which may lead to

automatic, mindless responses that lack genuine meaning. It also supports more comfortable interactions by reducing users' perceived obligation to follow social etiquette. Finally, it may prevent users from later feeling unintentionally deceived or experiencing disillusionment upon realizing that the voice was not human.

6.2 Responsive Vocal Characteristics

Our study emphasizes the importance of incorporating responsive and flexible paralinguistic cues, such as prosody, speech rate, response latency, and pauses, to support more fluid and rich voice interactions. To achieve this responsiveness, voice assistants should be capable of sensitively and subtly adjusting these cues in real time, in alignment with the user's conversational cues and expectations. For example, voice assistants could respond more quickly to short queries, adapt their speaking pace to match the user's speech—speaking more briskly when the user talks fast, and slowing down slightly when the user speaks more slowly. They could also pause or stop immediately when the user interrupts, and finely calibrate turn-final prosody to support smooth and predictable turn-taking. Once such seamless and effortless interactions become feasible, users will be able to complete tasks more quickly and efficiently through voice-based interfaces. In particular, enhanced responsiveness, alongside faster processing, could alleviate users' frustration and sense of being blocked in scenarios like drive-through voice-ordering services or customer service call centers. This, in turn, could reshape existing stereotypes about voice assistants as having limited conversational abilities. Additionally, as synthetic voices become increasingly natural, even slight mismatches in tone or prosody are more likely to create a sense of awkwardness or emotional disconnection. In response, voice assistants should convey vocal empathy and adjust their tone appropriately to reflect the user's emotional state and context. For instance, when a user expresses frustration or anger, a cheerful voice may need to be gently softened to better align with the user's mood. Rather than having voice assistants completely replicate human voices, we propose a voice design that selectively and carefully curates vocal characteristics to meaningfully enhance responsiveness to users' conversational cues and contexts, fostering smoother and more efficient interactions.

7 LIMITATIONS AND FUTURE WORK

In this section, we would like to point out some of the limitations regarding our study and the possible future work that could branch out from there. First, our study focuses on users' short and initial encounters with both artificial- and human-sounding voice assistants, with an average call duration of around one minute. Such a brief exposure limits the depth and complexity of interaction and may not fully observe how user experiences evolve over prolonged periods [18, 28, 47]. Future research could extend to longer interactions or even long-term studies, examining how vocal attributes influence user perceptions and behaviors over time.

Second, as indicated in our findings (Section 4.2), 9 out of 12 participants in the human-sounding condition concluded that the voice was human—either feeling confused about whether it was a system or never questioning its human identity. Although our study aimed to capture both the perception and experience of such ambiguity, this perception likely influenced the overall interaction and should be taken into account when interpreting the findings. Additionally, it is important to note that the human-sounding voice was, in fact, produced by a real human speaker (wizard), reflecting a future scenario in which the paralinguistic features of voice assistants could respond with fine-grained precision. Future research may build upon our findings to explore how user experiences differ when such systems become technically feasible.

Third, our study examined a transactional context, specifically scheduling interview times, which involved simple and straightforward tasks. However, in situations where social interactions play a significant role, such as counseling or handling complaints, further exploration may be needed. Future work could investigate not only transactional contexts (performing functions or tasks) but also social interactions with voice assistants to determine how voices should be designed and assigned to ensure engaging and meaningful communication.

Fourth, the diversity of human-sounding characteristics is vast, but our study only explored a subset of this richness, such as filled pauses, natural paralinguistic cues (subtle responsive prosody, speech rate, and latency), and personal accents. Excluded from our focus were overly emotional tones, disfluencies that impair intelligibility (e.g., repetition, false starts, and verbosity), and human noises (e.g., laughter, yawns, and breathing). Future studies could investigate a broader range of human-sounding attributes, particularly focusing on emotional aspects and balancing empathy with social awkwardness, to enhance our understanding of voice design.

Fifth, we utilized an artificial-sounding voice, primarily focusing on a consistent and intelligible female voice commonly used in earlier voice assistants. Based on our findings, we proposed creating machine-unique, transparent voices that convey identity solely through the voice alone. In this regard, Aylett et al. pioneered exploratory work on machine-unique voices—suggesting performance-style voices (e.g., actors playing a character) [1] and mixing speech synthesis with semantic-free utterances (e.g., the beeps, squeaks, and clicks found in Wall-E or R2-D2) [2]. This approach draws on the rich history of robotic sounds in science fiction films. Works by Disney and Spielberg can serve as valuable references for voice design. Future work could be inspired by how robot and agent characters sound in science fiction media.

Lastly, we predetermined the structure of the dialogue flow and scripts for scheduling tasks. This approach was intended to provide a comparable user flow for both human-sounding and artificial-sounding voices, while also ensuring error-free conversations and minimizing bias from error situations. Notably, our study found that users experienced awkwardness and a sense of disconnection when interacting with human-sounding voices over a laggy network. Regardless of how advanced the conversational abilities of large language model (LLM)-based voice assistant systems become, edge cases such as API call failures, network errors, hallucinations, and hyper-precision distortions are likely to occur. Future research examining user experiences with human-sounding voices in such unexpected error situations could provide valuable insights.

8 CONCLUSION

The study compared user perceptions and interactions with artificial-sounding and human-sounding voice assistants in a voice-only

phone call context embedded in everyday life. Our findings revealed seven themes that illustrate how specific conversational patterns differed between the two conditions and how participants socially perceived and responded to them. These findings led to a discussion on how the voices of conversational agents should be designed: to reveal their identity, foster natural and engaging interactions, and relieve social etiquette burdens while maintaining empathy. Furthermore, we proposed design implications that emphasize the importance of transparency and responsiveness in voice design. This study contributes to the evolving discourse by offering timely insights into designing voices for future voice assistants that blend artificial and human qualities. These insights may also inspire HCI researchers and designers to reimagine voice design while addressing associated ethical and usability challenges.

References

[1] 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. doi:10.1145/3290607.3310422

[2] 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. doi:10.1145/3371382.3378330

[3] Baidu Research. 2017. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. http://research.baidu.com/Blog/index-view?id=91. Accessed: 20 Apr. 2025.

[4] 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

[5] Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1994. Wizard of oz prototyping: When and How. *Proc. CCI Working Papers Cognit. Sci./HCI, Roskilde, Denmark* (1994).

[6] Paul Boersma and David Weenink. 2022. Praat: doing phonetics by computer (Version 6.2.06). http://www.praat.org

[7] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. 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, 1–11. doi:10.1145/3290605.3300270

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[9] 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. doi:10.1145/3313831.3376789

[10] 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. doi:10.1145/3359325

[11] 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: 20 Apr. 2025.

[12] 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. doi:10.1145/3313831.3376461

[13] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019a. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019a), 349–371. doi:10.1093/iwc/iwz016

[14] 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. 2019b. 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. doi:10.1145/3290605.3300705

[15] 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. doi:10.21437/Interspeech.2020-1336

[16] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can i Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. doi:10.1145/3098279.3098539

[17] 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. doi:10.1145/3544548.3581281

[18] Maartje MA de Graaf, Somaya Ben Allouch, and Jan AGM van Dijk. 2018. A phased framework for long-term user acceptance of interactive technology in domestic environments. *New Media & Society* 20, 7 (2018), 2582–2603. doi:10.1177/1461444817727264

[19] 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. doi:10.1145/3491102.3517564

[20] 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. doi:10.1145/3338286.3340116

[21] Laila Dybkjær, Niels Ole Bernsen, and Hans Dybkjær. 1993. Knowledge acquisition for a constrained speech system using WoZ. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Utrecht, The Netherlands. https://aclanthology.org/E93-1061

[22] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8 (2008), 630–645. doi:10.1016/j.specom.2008.04.002 Evaluating new methods and models for advanced speech-based interactive systems.

[23] Norman M. Fraser and G.Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (1991), 81–99. doi:10.1016/0885-2308(91)90019-M

[24] Chris Geison. 2019. what in the ux is "wizard of oz testing"? https://www.answerlab.com/insights/wizard-of-oz-testing. Accessed: 20 Apr. 2025.

[25] Google DeepMind. 2016. WaveNet: A generative model for raw audio. https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio. Accessed: 20 Apr. 2025.

[26] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634. doi:10.1016/j.csl.2010.10.003

[27] Elliott M Hoey and Kobin H Kendrick. 2017. Conversation analysis. *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (2017), 151–173.

[28] Evangelos Karapanos, John Zimmerman, Jodi Forlizzi, and Jean-Bernard Martens. 2009. User Experience over Time: An Initial Framework. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 729–738. doi:10.1145/1518701.1518814

[29] Reem Khokhar. 2022. It's perfectly normal to hate phone calls. https://lifestyle.livemint.com/smart-living/innovation/its-perfectly-normal-to-hate-phone-calls-111644903756365.html. Accessed: 20 Apr. 2025.

[30] 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. doi:10.1145/3532106.3533528

[31] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021b. 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. doi:10.1145/3411764.3445579

[32] 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). doi:10.3389/fnbot.2020.593732

[33] Brian Lathrop, Hua Cheng, Fuliang Weng, Rohit Mishra, Joyce Chen, Harry Bratt, Lawrence Cavedon, Carsten Bergmann, Tess Hand-Bender, Heather Pon-Barry, et al. 2004. A Wizard of Oz framework for collecting spoken human-computer dialogs: An experiment procedure for the design and testing of natural language in-vehicle technology systems. In *Proceedings of the 12th World Congress on Intelligent Transport Systems*.

[34] Yaniv Leviathan. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. https://blog.research.google/2018/05/duplex-ai-system-for-natural-conversation.html. Accessed: 20 Apr. 2025.

[35] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100. doi:10.1016/j.chb.2014.04.043

[36] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. doi:10.1145/2858036.2858288

[37] 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: 20 Apr. 2025.

[38] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 277–284. doi:10.1145/169059.169215

[39] 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: 20 Apr. 2025.

[40] Microsoft. 2024. Vall-E. https://www.microsoft.com/en-us/research/project/vall-e-x/. Accessed: 20 Apr. 2025.

[41] 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*.

[42] 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. doi:10.1016/S0747-5632(02)00081-X

[43] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3173574.3173580

[44] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. doi:10.1111/0022-4537.00153

[45] 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. doi:10.1145/191666.191703

[46] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics* 5 (2013), 171–191.

[47] William T. Odom, Abigail J. Sellen, Richard Banks, David S. Kirk, Tim Regan, Mark Selby, Jodi L. Forlizzi, and John Zimmerman. 2014. Designing for Slowness, Anticipation and Re-Visitation: A Long Term Field Study of the Photobox. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1961–1970. doi:10.1145/2556288.2557178

[48] James O'Donnell. 2024. OpenAI released its advanced voice mode to more people. Here's how to get it. https://www.technologyreview.com/2024/09/24/1104422/openai-released-its-advanced-voice-mode-to-more-people-heres-how-to-get-it/. Accessed: 20 Apr. 2025.

[49] Jeesun Oh, Hyeonjeong Im, and Sangsu Lee. 2024c. Toward a Third-Kind Voice for Conversational Agents in an Era of Blurring Boundaries Between Machine and Human Sounds. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 55, 7 pages. doi:10.1145/3640794.3665880

[50] 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. doi:10.1002/isaf.1443

[51] OpenAI. 2024. Navigating the Challenges and Opportunities of Synthetic Voices. https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices. Accessed: 20 Apr. 2025.

[52] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174214

[53] 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. doi:10.1177/14614448211024142

[54] 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). doi:10.3389/fpsyg.2022.787499

[55] 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. doi:10.1177/0956797617713798

[56] 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: 20 Apr. 2025.

[57] 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: 20 Apr. 2025.

[58] 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. doi:10.1145/3196709.3196772

[59] 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. doi:10.1016/j.chb.2019.04.001

[60] Nicole Shechtman and Leonard M. Horowitz. 2003. Media inequality in conversation: how people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 281–288. doi:10.1145/642611.642661

[61] 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.

[62] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. 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, 1–9. doi:10.1145/3290605.3300768

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

[64] 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. doi:10.1145/3290605.3300833

[65] Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a Toy": Exploring Older Adults' Reasons for Using, Limiting, and Abandoning Echo. 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. doi:10.1145/3313831.3376760

[66] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3173574.3173782

[67] Biao Wang. 2024. NotebookLM now lets you listen to a conversation about your sources. https://blog.google/technology/ai/notebooklm-audio-overviews/. Accessed: 20 Apr. 2025.

[68] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023a).

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

[70] Yuxuan Wang and RJ Skerry-Ryan. 2018. Expressive Speech Synthesis with Tacotron. https://ai.googleblog.com/2018/03/expressive-speech-synthesis-with.html. Accessed: 20 Apr. 2025.

[71] 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).

[72] 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. doi:10.1016/j.jesp.2014.01.005