frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 13 May 2022
doi: 10.3389/fpsyg.2022.787499

Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance

Simon Schreibelmayr and Martina Mara*

LIT Robopsychology Lab, Johannes Kepler University Linz, Linz, Austria

OPEN ACCESS

Edited by:
Jessica Martina Szczuka, University of Duisburg-Essen, Germany

Reviewed by:
Marcel Finkel, University of Duisburg-Essen, Germany
Franziska Gaiser, Leibniz-Institut für Wissensmedien (IWM), Germany
Katharina Kühne, University of Potsdam, Germany

****Correspondence:***
Simon Schreibelmayr
simon.schreibelmayr@jku.at

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *30 September 2021*
Accepted: *24 March 2022*
Published: *13 May 2022*

Citation:
Schreibelmayr S and Mara M (2022) Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. Front. Psychol. 13:787499. doi: 10.3389/fpsyg.2022.787499

The growing popularity of speech interfaces goes hand in hand with the creation of synthetic voices that sound ever more human. Previous research has been inconclusive about whether anthropomorphic design features of machines are more likely to be associated with positive user responses or, conversely, with uncanny experiences. To avoid detrimental effects of synthetic voice design, it is therefore crucial to explore what level of human realism human interactors prefer and whether their evaluations may vary across different domains of application. In a randomized laboratory experiment, 165 participants listened to one of five female-sounding robot voices, each with a different degree of human realism. We assessed how much participants anthropomorphized the voice (by subjective human-likeness ratings, a name-giving task and an imagination task), how pleasant and how eerie they found it, and to what extent they would accept its use in various domains. Additionally, participants completed Big Five personality measures and a tolerance of ambiguity scale. Our results indicate a positive relationship between human-likeness and user acceptance, with the most realistic sounding voice scoring highest in pleasantness and lowest in eeriness. Participants were also more likely to assign real human names to the voice (e.g., "Julia" instead of "T380") if it sounded more realistic. In terms of application context, participants overall indicated lower acceptance of the use of speech interfaces in social domains (care, companionship) than in others (e.g., information & navigation), though the most human-like voice was rated significantly more acceptable in social applications than the remaining four. While most personality factors did not prove influential, openness to experience was found to moderate the relationship between voice type and user acceptance such that individuals with higher openness scores rated the most human-like voice even more positively. Study results are discussed in the light of the presented theory and in relation to open research questions in the field of synthetic voice design.

Keywords: speech interface, voice assistant, human–robot interaction, synthetic voice, anthropomorphism, uncanny valley, application context, user acceptance

INTRODUCTION

Talking machines have found a place in our lives. They are supposed to assist us in a range of activities, be it performing an online search, navigating the way, or just letting us know when the spaghetti is ready. Around the world, 4.2 billion digital voice assistants, such as Amazon's Alexa or Apple's Siri, are already employed. By 2024, the number of digital voice assistants is predicted to reach 8.4 billion units, a number greater than the world's human population (Juniper, 2019; Statista, 2021). Over the upcoming years, it is thus clear that ever more people will use spoken language to interact with machines—and these machines will eventually sound more and more human-like (Meinecke, 2019; Statista, 2021). Google Duplex, to mention one of the more recent innovations in the field of speech synthesis, gives us a glimpse of the future where computer voices might actually be indistinguishable from real people (Oord et al., 2016; Google Duplex, 2018). However, unlike us humans, who cannot fundamentally change the sound of our voices except for slight adaptations to the situation and interlocutor, synthetic voices are "design material" (Sutton et al., 2019) allowing for customization (Amazon, 2017; Polly, 2019; Cohn and Zellou, 2020). Depending on deliberate design decisions, computer-generated voices may thus sound more female or male, younger or old, more bored or excited—more human or mechanical.

Since virtually no new skills need to be learned for natural language communication with computers and speech interfaces are therefore considered particularly intuitive even for non-experts (e.g., Nass and Brave, 2005), synthetic voices are being used in a growing number of technological products. Besides voice assistants, these include conversational agents, customer service bots, navigation systems, social robots, vending machines, or even AI therapists (Niculescu et al., 2013; Chang et al., 2020). As voice interfaces evolve and their areas of application continue to expand, it must be ensured that the needs of users are adequately addressed. If important acceptance factors are not accounted for in their design, this may not only backfire economically, but also have negative consequences for the psychological wellbeing of users. User-centered research is therefore needed to gain a better understanding of effects of vocal human-likeness in machines and to investigate what types of synthetic voices are considered acceptable in different contexts of use.

To date, we know only little about whether realistically human-sounding computer voices would elicit particularly positive or negative user responses, and if it matters whether we think of a more social application such as a talking care robot or a more formal one such as a financial assistant. In a recent attempt to shed light on this matter, Kühne et al. (2020) found, contrary to their expectations, that participants generally liked highly human-like computer voices more than synthetically sounding ones. Against the background of the popular Uncanny Valley hypothesis (Mori, 1970) and empirical findings on visual or behavioral human-likeness in robots (Bartneck et al., 2007; Mara and Appel, 2015a,b; Appel et al., 2016; Mathur and Reichling, 2016), however, it could be assumed that a too realistic imitation of the human would lead to aversive responses.

Given the mixed perspectives in the literature, the rapidly advancing progress in the development of human-sounding synthetic voices, and the diverse purposes for which speech interfaces may be used in society, controlled user studies are required that include a range of more or less human-like voices while also considering contextual and individual differences. This is where the present work comes in with fourfold objectives. Based on a lab experiment with five different voices, supposedly belonging to a service robot, it shall contribute to answering the following questions:

(RQ1) Voice realism and anthropomorphism:
Are machines with more realistic voices actually more anthropomorphized than machines with less realistic voices?

(RQ2) Human-likeness and the Uncanny Valley:
Is the degree of perceived human-likeness related to how eerie or pleasant users evaluate a given voice?

(RQ3) Application context and acceptance of vocal human-likeness:
Does the acceptance of vocal human-likeness depend on the assumed application context, and more specifically on whether it is a social context?

(RQ4) User personality and acceptance of vocal human-likeness:
Considering tolerance of ambiguity and the Big Five personality factors, do individuals differ in how positively they evaluate vocal human-likeness?

Before we describe the conducted experiment in more detail, the underlying theoretical and empirical literature is presented in the following sections. For better comprehensibility, hypotheses are laid out directly below the literature section they were derived from.

Human-Like Voice as Anthropomorphic Cue

The human voice is the most impactful sound in our lives. It represents a very important component of interpersonal communication and it transmits highly relevant information about its creator (Kaplan et al., 1995; McGee et al., 2001). The moment we start to speak, we automatically reveal information about our biological, psychological, and social status. Research has demonstrated that characteristics, such as a person's gender, age, affect, and their membership in social or ethnic groups, can be inferred from the voice only, even if the person was previously unknown to the judge (Giles et al., 1979; Eagly and Wood, 1982; Kohlberg et al., 1987; Krauss et al., 2002; Pinker, 2003; Tiwari and Tiwari, 2012; Smith et al., 2016).

Looking at the crucial role of human voice to exchange information and to interpret others in our social life, it is not surprising that voice emitted by a computer is considered a particularly strong anthropomorphic cue (Nass and Brave, 2005; Qiu and Benbasat, 2009; Eyssel et al., 2012; Whang and Im, 2021),

along with visual cues, such as human-like embodiment or non-verbal behavior of a machine (cf. Mara and Appel, 2015a,b). Anthropomorphism describes the widespread tendency to attribute human characteristics, motivations, intentions, or emotions to non-human entities, or in short, to sense something human where there is actually nothing human (Epley et al., 2007). This can happen with things that do not use natural speech or resemble human appearance at all, such as cuddly toys or even plants. According to Theory of Anthropomorphism of Epley et al. (2007), however, readily observable human-like features increase an object's likelihood of being anthropomorphized because they facilitate the accessibility of anthropocentric knowledge structures and thus increase the chance that such knowledge will be applied to the non-human target. This is in line with Nass and colleagues' Computers Are Social Actors paradigm (CASA, Nass et al., 1994; Reeves and Nass, 1996; Nass and Brave, 2005), which posits that individuals mindlessly apply social heuristics from interpersonal interactions to their interactions with computers. According to the authors, perceiving a computer as social actor is particularly likely when it takes on a role that was typically fulfilled by a human (e.g., tutor, salesperson, and therapist), when it is interactive, or when it uses natural speech (Nass et al., 1994; Nass and Brave, 2005).

In support of these theories, empirical research has found, for example, that consumers perceive voice assistants as independent agents detached from the company behind them (Whang and Im, 2021), that different voices emitted by the same computer are treated as distinct social actors (Nass et al., 1994), that the use of voice in online questionnaires elicits socially desirable responses comparable to the way a real human interviewer would (Couper et al., 2001; Tourangeau et al., 2003), and that people deduce personality cues from synthetic voice (Nass and Lee, 2001). Furthermore, initial evidence suggests that it is not just the use of voice *per se* that matters, but that greater anthropomorphization occurs with more natural computer voices than with less natural ones (Eyssel et al., 2012; Ilves and Surakka, 2013; Baird et al., 2018).

Various validated self-report scales exist to measure how much human someone sees in a machine (Bartneck et al., 2009; Ho and MacDorman, 2010; Carpinella et al., 2017). Besides, a common expression of anthropomorphism in everyday life (and also a common strategy in product marketing) is giving a human name to an object (Epley et al., 2007). Name-giving and anthropomorphism have been previously associated in the scientific literature. For example, human first names have been used to experimentally manipulate the perceived human-likeness of a machine (e.g., Qiu and Benbasat, 2009; Waytz et al., 2010). Recently, Brédart (2021) studied this relation from the flip side and revealed that people with higher anthropomorphic tendencies were also more likely to call personal objects by a proper name. While we found no existing studies on the relationship between strength of anthropomorphism and name-giving with respect to synthetic speech, there is evidence that, depending on the perceived human-likeness of a computer voice, individuals also imagine the embodiment behind the voice to be more or less human (e.g., with or without human face, hair, and hands; Mara et al., 2020), which may also reflect anthropomorphism.

From the literature presented, we derive the following initial hypotheses regarding the relationship between voice realism and anthropomorphic attributions:

> *H1a*: The more realistic a voice sounds, the more *human-like* it is rated.
> *H1b*: The more realistic a voice sounds, the more likely participants assign a real *human name* to the talking robot in a name-giving task.
> *H1c*: The more realistic a voice sounds, the more likely participants describe the talking robot to have a *human-like appearance* in an imagination task.

User Evaluations of Human-Like Machines: Pleasant or Uncanny?

Manufacturers of tech gadgets in many cases seek to fuel user perceptions of their products as human-like. In the context of this paper, voice assistance systems that often have not only human names but also specially created backstories (West et al., 2019), are the best example of how companies assume anthropomorphism to be associated with positive customer opinions.

Consistent with this popular belief, findings from a few recent studies indeed indicate more favorable user evaluations for greater human-likeness in computer voices. Kühne et al. (2020) drew a comparison between two currently available synthetic female voices (CereVoice, IBM Watson) and a real woman's voice. Results indicate that the real human voice was rated as most pleasant, intelligible, likable, and trustworthy. Anecdotal evidence from two other exploratory studies suggests similar patterns Baird et al. (2018) asked 25 listeners to evaluate the likability and human-likeness of 13 synthesized male voices and found likability to increase consistently with human-likeness. Based on data from 30 listeners, also Romportl (2014) reported that most though not all participants preferred a more natural female voice over an artificial sounding one. These results are also in line with two recent meta-analyses that overall show beneficial effects of—here, mostly visual—anthropomorphic design features for embodied robots and chatbots (e.g., on affect, attitudes, trust, or intention to use), although the dependence of these effects on various moderators (e.g., robot type, task type, and field of application) points to more complex relationships between human-likeness and user responses (Blut et al., 2021; Roesler et al., 2021).

The literature, however, also features a number of studies that report non-favorable user reactions to high levels of human-likeness in machines. For example, in several experiments from the field of human-robot interaction it was found that people prefer more machine-like robot appearances over more human-like ones (Bartneck et al., 2007; Broadbent et al., 2011; Mara and Appel, 2015a,b; Mathur and Reichling, 2016; Vlachos et al., 2016; Jia et al., 2021). Works that suggest negative effects of anthropomorphic designs typically refer to the Uncanny Valley hypothesis (Mori, 1970; Mori et al., 2012), which proposes

a non-linear relationship between the human-likeness of an artificial character and the valence it evokes in its observers. According to Mori's hypothesis, in a generally low range of human-likeness, pleasantness grows with increasing realism. At a point of rather high human-likeness, however, the effect reverses and the artificial entity is perceived as eerie or threatening. Only when the entity's degree of realism reaches near-perfection or perfection will pleasantness go up again, since no distinction can be made any longer between artificial and human (Mori et al., 2012; Mara et al., 2022). Various perceptual and cognitive mechanisms have been suggested to underlie uncanny experiences (cf. Diel and MacDorman, 2021). These include categorical uncertainty or prediction difficulties if features of a given entity seem to belong to different conceptual categories (e.g., a mechanoid robot head with a human-like voice, Mitchell et al., 2011; Meah and Moore, 2014).

In summary, given some recent empirical findings on synthetic speech, it could be assumed that voices that are perceived as more human-like are also perceived as more pleasant and less eerie (Romportl, 2014; Baird et al., 2018; Kühne et al., 2020). Against the background of the Uncanny Valley phenomenon, however, expectations would go in a different direction: On the one hand, it could be assumed that highly realistically sounding voices are evaluated as eerier and less pleasant than either a perfect imitation of the human voice or mechanically sounding voices. This would depict the curvilinear relationship between human-likeness and elicited valence as originally predicted by Mori (1970). On the other hand, if we refer to conflicting cues and categorical uncertainty as potential mechanisms behind uncanny experiences (cf. Burleigh et al., 2013; Diel and MacDorman, 2021), a mismatch between the sound of a voice (e.g., highly human-like) and available information about the speaker (e.g., "It is a robot") could also be assumed to trigger eeriness. Since we consistently introduce each of the five voices in our study as a "robot voice," following this idea, the real human voice might be perceived as the greatest mismatch and therefore possibly evokes greatest eeriness. Overall, given the various plausible assumptions that could be deduced from the theoretical and empirical literature, we remain with non-directional hypotheses on the relationship between voice realism, pleasantness, and eeriness at this point:

> *H2a*: *Eeriness* evaluations differ between the voices and their human-likeness ratings.
> *H2b*: *Pleasantness* evaluations differ between the voices and their human-likeness ratings.

Acceptance and Application Context

Computer voices are supposed to find use in a wide variety of applications, from care or companion robots (Bendel, 2022) to AI-based financial assistants (Kaur et al., 2020). While there is hardly any research on the contextual acceptance of voice interfaces to date, recent meta-analyses from the broader field of human–robot interaction suggest that user acceptance is unlikely to be independent of the application area and the tasks for which a robot is to be used (Blut et al., 2021; Roesler et al., 2021). For example, Ullman et al. (2021) show in a series of studies that robots are consistently regarded as less trustworthy in social application contexts than in non-social ones. This is in line with an experiment, which saw the robot iCub being trusted more for functional tasks, such as image analysis than for social tasks (Gaudiello et al., 2016). Transnational surveys from Europe also indicate that many people are generally more positive about the use of robots in areas, such as space exploration or manufacturing than in areas that typically require social–communicative skills and empathy, with only 3–4% of Europeans welcoming a priority use of robots for the care of children or the elderly (Eurobarometer, 2012).

Since different application areas raise different expectations about what a machine must be able to do, it seems reasonable to assume that the degree of human-likeness considered appropriate and acceptable by users is also context-dependent. A few empirical studies have so far addressed potential interaction effects of anthropomorphism and application context. In Roesler and colleagues' recent experiment (Roesler et al., 2022), participants had to choose one out of various robot pictures that differed in visual human-likeness based on different context descriptions. A lower degree of human-likeness was found to be preferred for industrial application and a higher degree of human-likeness for social application, while there were no clear preferences in the service domain. This is consistent with a previous study (Goetz et al., 2003), which also observed a preference for human-like robots for social tasks, but machine-like robots for investigative tasks. Oyedele et al. (2007) found tentative evidence for an interaction effect in that more human-like robots were assessed more positively in an imagined household context, while the degree of human-likeness was irrelevant for acceptance in other contexts. In contrast, results by Jung and Cho (2018) indicate no interaction as images of highly human-like robots were rated more negatively than mechanoid robots across several contexts.

Taken together, empirical findings seem to suggest that while overall acceptance for the use of robots in social application domains is lower than for non-social domains, acceptance within social applications increases with the degree to which a machine is perceived human-like. Following definitions from Social Robotics, for the purpose of this study, social applications are defined as ones in which machines act as "social partners" (Mejia and Kajikawa, 2017), engage in meaningful two-way interactions, build emotional resonance, understand human states, and respond to them according to social rules (Duffy, 2003; De Graaf et al., 2015). This was described to be the case with robots meant to provide caregiving or companionship, among others (Mejia and Kajikawa, 2017).

With respect to context-dependent differences in the acceptance of computer voices, we derive the following hypotheses from the literature:

> *H3a*: Independent from voice type, *acceptance* for the use of voice interfaces is lower for social applications (care, companionship) than for non-social applications (business & finance, information & navigation).

H3b: The more realistic a voice sounds and the more human-like it is perceived, the more likely it is to be *accepted* for use in *social application* areas (care, companionship).

Acceptance and User Personality

Taking personality psychological approaches into account, it can be assumed that the evaluation and acceptance (or rejection) of anthropomorphic machines is not only determined by design parameters of the machine itself and its application area, but also by user-specific factors. Two of the personality traits of the famous five-factor model (FFM or "Big Five," Digman, 1990; John et al., 1991), namely, openness to experience and neuroticism, have been associated with the acceptance of new technologies in many studies.

Openness to experience, that is, a person's tendency to prefer novelty over routine and to have a broad rather than a narrow range of interests, has been found to correlate, among others, with more positive attitudes toward robots (Morsunbul, 2019), acceptance of robots (Esterwood et al., 2021), acceptance of autonomous vehicles (Gambino and Sundar, 2019; Zhang et al., 2020), and with personal innovativeness in IT (Nov and Ye, 2008). In a study on a new teleworking software (Devaraj et al., 2008), openness turned out to be the only of the "Big Five" personality factors that had a direct impact on intentions to use beyond the two core predictors (usefulness, ease of use) of the widely used Technology Acceptance Model (TAM, Davis, 1989). Furthermore, people with higher openness scores were found to be less prone to technophobia (Anthony et al., 2000; Maricutoiu, 2014).

In contrast, individuals with higher neuroticism scores, that is, those who are more likely to experience emotional instability, negativity, anxiety, and irritation, showed less eagerness to adopt new technologies (e.g., Charness et al., 2018; Zhang et al., 2020) and were found to suffer more often from technophobia (Maricutoiu, 2014). Persons who scored higher in neuroticism also experienced highly human-like robots as eerier and less warm in a study (MacDorman and Entezari, 2015), which could be interpreted as a greater uncanny valley sensitivity.

Apart from the "Big Five," initial empirical evidence indicates that persons who generally respond negatively to ambiguous stimuli or who are sensitive to a lack of structure describe highly human-like machines as eerier than others (Lischetzke et al., 2017). If a categorization process is hindered, for example due to machine characteristics that are close to categorical boundaries or due to conflicting cues (a robot as per information, but with a very natural voice), it could thus be assumed that people who score low on tolerance of ambiguity may experience discomfort or even uncanniness (cf. Bochner, 1965; Norton, 1975; Freeston et al., 1994; Furnham and Ribchester, 1995; Robinson et al., 2003; Robinson, 2004; Oshio, 2009; MacDorman and Entezari, 2015).

Based on the literature presented, we consider individual differences to play a role in user responses to human-like computer voices. Following findings from technology acceptance studies and the Uncanny Valley literature, we assume neuroticism and low tolerance of ambiguity to add to higher eeriness ratings of human-like voices, whereas greater openness to experience should add to greater acceptance for applying human-like computer voices, as reflected by the following hypotheses:

H4a: The relationship between perceived *human-likeness* and *eeriness* of a voice is moderated by participants' *tolerance of ambiguity*.
H4b: The relationship between perceived *human-likeness* and *eeriness* of a voice is moderated by participants' *neuroticism*.
H4c: Differences in user *acceptance* between the voices are moderated by participants' *openness to experience*.

MATERIALS AND METHODS

To test our assumptions, we compared user responses to speech recordings of a total of five female-sounding voices supposed to belong to a (not visible) service robot in a randomized controlled lab experiment with constant listening conditions. In the following, we give a detailed description of the voice stimuli created for this study, the characteristics of our sample, the study procedure, and the measures used.

Voice Stimuli

Recordings of five different voices (*human*, *synthetic I*, *synthetic II*, *metallic*, *comic*) were created as auditory stimuli. All speech samples were in German. Duration, speech content, and voice gender (female) were held constant to control for potential confounding effects. The total length of each recording was 2 min and 20 s and consisted of 306 words. The speech content represented an introduction of the history and technical functionality of robots. It was written with the intent (i) to be thematically apt but relatively neutral, (ii) not to bias the participants' acceptance of specific robot application areas, and (iii) not to encourage anthropomorphic inferences which may systematically impact the perception of certain voice types in different ways than others (Fink et al., 2012).

In order to cover a wide range of varying vocal realism across our stimuli, recordings of a real person, professional synthetic voices as well as less realistic sounding modifications of synthetic voices were included (see **Table 1**). Subsequently, an overview of the five experimental voices is given.

Human

This speech sample was recorded by a professional voice-trained speaker in a quiet room using the recording software "Logic" and a large-diaphragm condenser microphone with a cardioid characteristic called "Rode NT-1 A." As the participants were supposed to believe that this real human voice was also artificially generated, noises like exhaling and inhaling between the words were removed using the software "Adobe Audition" (Adobe Audition, 2019). This ensured that the voice sounded highly realistic yet not perfectly natural.

TABLE 1 | Description of the five experimental robot voices.

	Voice name	Speech engine	Modification
Real human	Human	(Pro speaker)	Breath sounds filtered
High human-likeness	Synthetic I	Amazon Polly (German)	Original version
	Synthetic II	Microsoft Hedda (German)	Original version
Low human-likeness	Metallic	Amazon Polly	Metallic effect, Echo (10%)
	Comic	Amazon Polly	Pitch shift (1.35)

Synthetic I

In this condition, the high-quality synthetic voice "Vicki" from Amazon Polly's text-to-speech portfolio (Polly, 2019) was used. Amazon described "Vicki" as a "voice of a similar fluency and naturalness as the German voice of Alexa" (Amazon, 2017).

Synthetic II

The voice "Hedda" represents an older text-to-speech system available on the Microsoft Speech Platform (Hedda, 2019). In comparison with synthetic I, this voice is more easily classified as artificial because of typically synthetic accentuations.

Metallic

Aiming for reduced vocal realism, here the original voice *synthetic I* was manipulated by means of a metallic echo effect (find details in **Appendix A**).

Comic

For this condition, the pitch of the original voice *synthetic I* was raised with the help of the software Voxal (2019) so that the voice sounded higher and more like a cartoon character (find details in **Appendix A**).

All recordings were cleaned with a manually created noise-removal filter using the software "Audacity" and adjusted to the same volume by normalizing the amplitude using the extension "dpMeter4" by "Audiveris" (Audacity, 2019; Audiveris, 2019; find details in **Appendix A**).

Sample Size Justification and Participants

The sample size required for the present between-subject experiment was calculated by a power analysis using G*Power (Cohen, 1992a; Faul et al., 2007). For the calculation, a medium effect size of $f = 0.30$ was assumed and α error probability was set to 0.05. In order to achieve a power $(1-\beta)$ of 85%, the analysis resulted in a recommended sample size of at least $N = 154$ to run an ANOVA. A total of 165 German-speaking individuals took part in our lab experiment. The participants were recruited at the campus of the Johannes Kepler University in Linz, Austria and through a snowball approach.[1] Data of two participants had to be excluded, because they reported not having responded conscientiously to all questions. Thus, the final sample consisted of 163 individuals (99 women, 64 men, no person of another or unknown gender identity), aged between 16 and 74 years ($M = 26.39$, SD = 9.64). Most of them were students (64.4%). 21.5% of participants stated they currently used a voice assistance system, such as Siri or Alexa, and 20.9% had personal experience with a robot at their home (e.g., lawn mower robot and vacuum cleaner robot). Their mean self-reported technology affinity (measured with a 5-point scale from 1 = low to 5 = high) was $M = 3.64$, SD = 1.21, overall indicating a slightly above-average interest in technology in our sample.

[1]Individuals who had already participated were asked to invite new study participants. A general introductory text about the study was provided to help recruiting new participants. Persons who had already participated in the experiment were sensitized to not communicate any additional information about the contents of the study to newly recruited persons.

Procedure

After arriving at the university's computer lab, participants received a short introduction by the experimenter, signed a consent form, and took a seat at one of the computers. They put on high-quality over-ear headphones (Beyerdynamic DT990 Pro) and started the experiment by clicking on the computer screen. At the same time, each person was automatically assigned to one of the five voice conditions ($N_{\text{Human}} = 34$, $N_{\text{Synthetic I}} = 34$, $N_{\text{Synthetic II}} = 33$, $N_{\text{Metallic}} = 31$, $N_{\text{Comic}} = 31$). The experiment began by asking participants to provide demographic information (including age, gender, and level of education) and to fill in personality questionnaires (including Big Five traits and tolerance of ambiguity). Next, they were told that they would now hear the first part of a voice recording of a new service robot, in which they would learn about the history and technical features of robots. This initial voice recording was 1 min 20 s long. No visual stimuli were presented while participants listened to one of the voices. After the first part of the recording, participants were asked to evaluate how pleasant, human-like and eerie they found the robot voice. Subsequently, the second half of the stimulus recording with a length of 1 min was played to them, again with the same voice variant as before. In the last part of the experiment, participants rated the degree of realism of the voice and indicated how much they would accept its use in different areas of application. In addition, participants were asked to physically envision the robot they had listened to, freely describe its appearance with a few keywords, and write down an appropriate name for it. Finally, some check items were queried (e.g., answered conscientiously and quality of headphones). The entire study was conducted by use of the software Questback (2018). The experiment took about 25 min per person. Participants were fully debriefed about the research background at the end of the experimental session. No financial compensation was provided for study participation.

Measures

Dependent Variables

We examined anthropomorphic attributions, eeriness, pleasantness, and acceptance as our dependent variables. The variable perceived realism was used as manipulation check (on a 9-point Likert scale).

Anthropomorphic Attributions

The perceived *human-likeness* of the speaking robot was assessed with five items on a five-point semantic differential scale (e.g., 1 = *synthetic*, 5 = *real*; 1 = *mechanical*, 5 = *organic*, adapted from Ho and MacDorman, 2010), which yielded an excellent reliability with Cronbach's $\alpha = 0.916$.

Assigned Name. In an open text box, participants provided a name for the robot that they felt was fitting to the robot they had listened to.

Imagined Embodiment. In a second open text box, participants described how they imagined the physical appearance of the robot they had listened to.

Eeriness and Pleasantness

Eeriness was measured with three items on a five-point semantic differential scale (e.g., 1 = *scary*, 5 = *comforting*, as example of an inverse coded item, adapted from Ho and MacDorman, 2010, Cronbach's $\alpha = 0.765$). The German items differed slightly from the English original items in favor of better comprehensibility (see **Table 2, Appendix B**).

Pleasantness was assessed by use of a single-item measure ("How pleasant did you find the voice?," ranging from 1 = *not at all* to 5 = *very much*).

Acceptance

Context-specific acceptance was measured with the help of one item for each application context ("How much would you agree with the use of the robot you listened to in the following areas?,"—Care,—Companionship,—Information & navigation,—Business & finance;—Entertainment,—Customer service, each ranging from 1 = *not at all* to 5 = *very much*).

With this selection of listed application contexts, we attempted to cover domains (a) that have also been included in previous studies, and (b) in which voice-enabled robots or AI systems are already in use today or are expected to be increasingly used in the upcoming years (e.g., Wada et al., 2003; Wada and Shibata, 2006; Eurobarometer, 2012; Aaltonen et al., 2017; Pérula-Martínez et al., 2017; Lopatovska et al., 2019). Following our definition in chapter 1.3, the domains "care" and "companionship" were classified as social applications, while "business & finance" and "information & navigation," where machines are usually not required to build emotional resonance or act as "social partners," were classified as non-social applications in the context of our paper. "Entertainment" and "customer service" were included for exploratory purposes.

To compare the *cross-context acceptance* between the voices, a mean score for each voice was built by averaging the acceptance scores across all contexts.

For the *context-specific acceptance index* (including all voices), a score was created by averaging across all voices to one acceptance score for each context.

Moderator Variables

Big Five Personality Dimensions

To assess personality factors, we used a 15-item short-scale from the Socio-Economic Panel (SOEP; see Schupp and Gerlitz, 2014), based on the Big Five Inventory by John et al. (1991) and Costa and McCrae (1985). Each personality dimension is determined by three items in this scale. Internal consistencies were moderate to good (*Openness to experience:* Cronbach's $\alpha = 0.73$, *Conscientiousness:* Cronbach's $\alpha = 0.64$, *Extraversion:* Cronbach's $\alpha = 0.80$, *Agreeableness:* Cronbach's $\alpha = 0.59$, *Neuroticism:* Cronbach's $\alpha = 0.70$). While we had formulated hypotheses regarding the role of *openness to experience* and *neuroticism*, the other Big Five variables were included for exploratory purposes.

Tolerance of Ambiguity

To measure the participants' tolerance of ambiguity we used 10 items assembled through a factor analysis by Radant and Dalbert (2003). The selection of the items is based on the 16-item short-scale developed by Schlink and Walther (2007). The scale showed a good internal consistency (Cronbach's $\alpha = 0.78$).

Manipulation Check

Realism was used as a manipulation check and assessed by use of a single-item measure ("How realistic does the voice of the robot sound in your opinion?," ranging from 1 = *not at all realistic* to 9 = *very realistic*).

RESULTS

Before testing our hypotheses, we examined if prerequisites of parametric analyses (normal distribution, homoscedasticity of the variances) were met by our data. As this was not the case for several variables, we decided to apply non-parametric test procedures (Kruskal–Wallis tests, Spearman's rank correlation). Significant differences in the *realism* ratings of the five voices indicate that our experimental manipulation worked [$H(4) = 56.491$, $p < 0.001$]. The real human voice was rated most realistic, the professional synthetic voices Synthetic I (by Amazon) and Synthetic II (by Microsoft) were ranked middle, and the modified synthetic voices were rated least realistic.

Voice Realism and Anthropomorphism

We hypothesized that the five voices would be anthropomorphized to varying degrees. Along with increasing levels of voice realism, participants were expected to more likely rate a voice as human-like (*H1a*), give it a real human name (*H1b*), and imagine the (invisible) speaking robot to have a human-like physical appearance (*H1c*).

In terms of human-likeness ratings, significant group differences between the five voices were found [*human-likeness:* $H(4) = 77.968$, $p < 0.001$; see **Table 2**], whereby the voice *Human* is distinct from all other voices in perceived human-likeness. The highest effect size (Cohen, 1992b) is $r = 0.96$ and corresponds to a strong effect describing the difference in human-likeness between the voice *Human* ($M = 3.85$, SD = 0.93) vs. *Metallic* ($M = 1.52$, SD = 0.42). Find all pairwise group comparisons in **Table 4 in Appendix C**. In **Figure 1**, voices are ranked in the order of their perceived human-likeness.

For the analysis of assigned names, the collected names were manually classified into five categories, which we created *post-hoc* on the basis of a first check of participant responses (1 = "*female real name*," 2 = "*male real name*," 3 = "*existent voice assistant*," 4 = "*fictional character*," 5 = "*mechanical*," $N = 158$; 5 missing). Two independent raters assigned each name to one of the classes. If they did not agree (in less than 5% of the cases), a collaborative decision was made.

A chi-square goodness-of-fit test revealed significant overall differences in the distribution of name classes, $X^2(4) = 117.316$, $p < 0.001$. As can be seen in **Figure 2** and **Table 5 (see Appendix C)**, nearly half (45.4%) of the names that participants came up with were real female first names (e.g., "Barbara" and "Julia"), whereas about a third (33.1%) were mechanical names (e.g., "T380" and "R-74"), 7.4% were real male first names (e.g., "Robert" and "Antonius"), 6.1% fictional character names (e.g., "C3PO" and "iRobot"), and 4.9% existing speech assistants' names (e.g., "Siri" and "Cortana").

To test *H1b*, a chi-square test including Monte Carlo Simulation (because of insufficient cell numbers <5; Hope, 1968; Sprent, 2007) was used. As expected, significant differences were found in the distribution of chosen names between the voices, $X^2(16) = 32.360$, $p = 0.007$, with the highest percentage of real human names (female/male first names) assigned to the voices Human and Synthetic I, whereas the lowest percentage of real human names was found for the voice Comic.

To test *H1c*, four independent evaluators rated the verbal descriptions of the robot's imagined physical embodiments *post-hoc* by means of a five-point Likert scale ranging from 1 = *very mechanical embodiment* to 5 = *very human-like embodiment*. A moderate inter-rater agreement was given (Fleiss' kappa $\kappa = 0.47$; Landis and Koch, 1977). After there were a couple of missing values in the embodiment descriptions, for the following group comparisons, the voices Human and Synthetic I were combined into a high vocal realism group, whereas the remaining voices Synthetic II, Comic, and Metallic were combined into a low vocal realism group. In line with our assumptions, a non-parametric Mann–Whitney *U*-test showed significant differences, indicating that the robot appearances were described as significantly more human-like after listening to one of the high vocal realism voices (Mdn = 3.5) than after listening to one of the low vocal realism voices (Mdn = 2.5), $U = 2006.50$, $Z = -2.99$, $p = 0.003$. Descriptions of robot appearances in the high vocal realism group included "Modelled after a female; friendly facial features and human-like behavior; blinking, head movements, female terminator?" or "female, white/light skin, blue eyes, young, cold." Exemplary descriptions from the low vocal realism group included "Metal and plastic case, screen with text, nothing human" or "a round white disc (…); simple modern design, smooth surface."

Human-Likeness and the Uncanny Valley

Next, we examined our assumptions regarding the relationship between vocal human-likeness and pleasantness as well as eeriness evaluations. Our non-directional hypotheses inferred that there would be significant group differences between the voices in both their eeriness scores (*H2a*) and their pleasantness scores (*H2b*).

As expected, significant group differences between the five voices were found both for eeriness [$H(4) = 48.468$, $p < 0.001$] and for pleasantness [$H(4) = 65.432$, $p < 0.001$; See **Figure 1, Table 2**]. As shown in **Table 6 (see Appendix C)**, across all voices, zero-order correlations indicate that human-likeness is negatively associated with the *eeriness* of a voice, $r_s(161) = -0.565$, $p < 0.01$, but strongly positively associated with *pleasantness*, $r_s(161) = 0.699$, $p < 0.01$. The real human voice was perceived as most human-like, but least eerie. *Pleasantness* and *eeriness* show a strong negative correlation, $r_s(161) = -0.666$, $p < 0.01$. Find all significant correlations across voices as well as for each voice separately in **Table 6 (see Appendix C)**.

After performing the Kruskal–Wallis tests, pairwise *post-hoc* comparisons were carried out for further analyses (all *p*s Dunn–Bonferroni adjusted). As shown in **Table 4 (Appendix C)**, 5 of 10 pairwise comparisons indicate significant differences in perceived eeriness and 6 of 10 in perceived pleasantness. The greatest effect for *eeriness* with $r = 0.70$ appears in the difference between the voices *Human* vs. *Metallic*. For *pleasantness*, the greatest effect of $r = 0.89$ was found for the difference between the voices *Human* vs. *Comic*.

TABLE 2 | Means and standard deviations of the ratings of the five voices.

Human-likeness*			Eeriness*			Pleasantness**		
	Mean	SD		Mean	SD		Mean	SD
All voices	2.27	1.13	All voices	2.81	0.93	All voices	2.81	1.20
Human	3.85	0.93	Human	2.14	0.80	Human	4.06	0.89
Synthetic I	2.19	0.60	Synthetic I	2.41	0.80	Synthetic I	3.15	0.96
Synthetic II	1.99	0.97	Synthetic II	2.82	0.79	Synthetic II	2.64	1.03
Comic	1.65	0.65	Comic	3.32	0.75	Comic	1.90	0.91
Metallic	1.52	0.42	Metallic	3.45	0.79	Metallic	2.16	0.87

*$N_{All} = 163$, $N_{Human} = 34$, $N_{Synthetic\ I} = 34$, $N_{Synthetic\ II} = 33$, $N_{Metallic} = 31$, $N_{Comic} = 31$. *Rated on a five-point semantic differential scale. **Rated on a five-point Likert scale from 1 (very unpleasant) to 5 (very pleasant).*

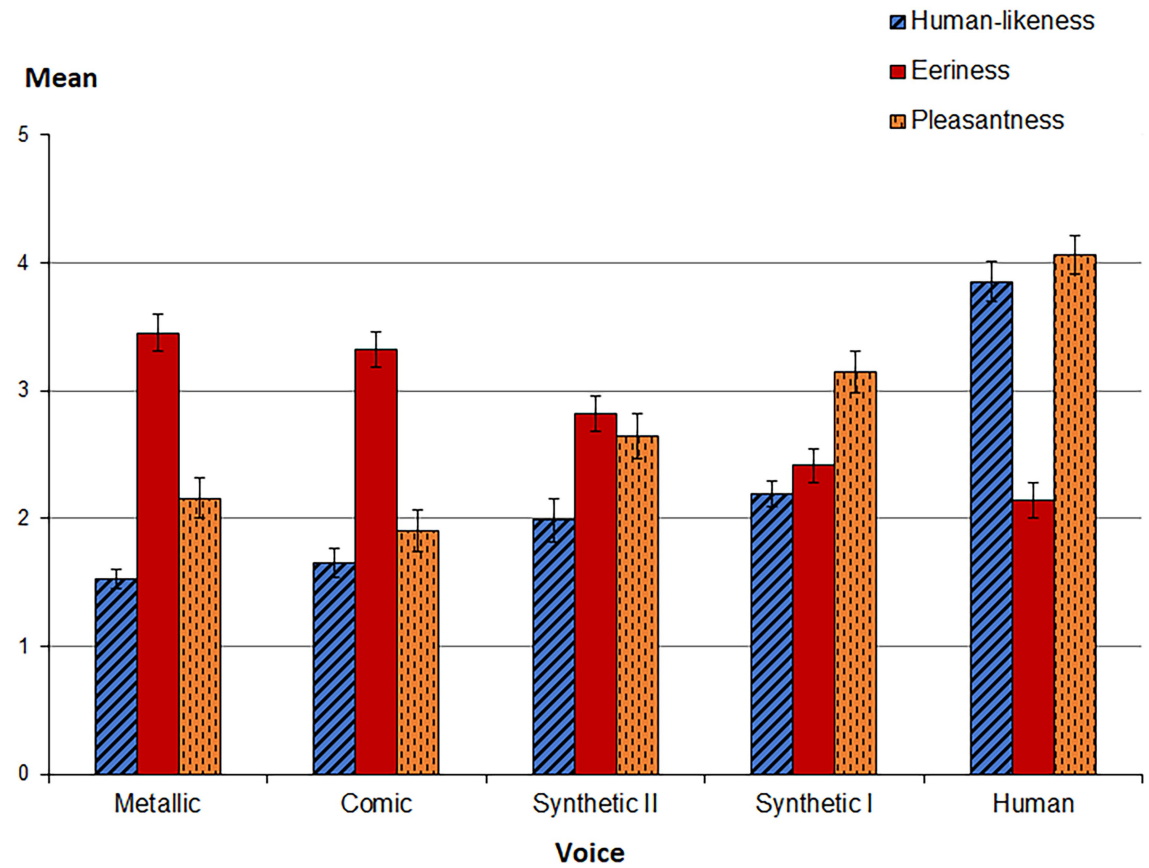

FIGURE 1 | The bar chart shows the mean values of the variables *Human-likeness*, *Eeriness,* and *Pleasantness* depending on the heard voice. The five voices are arranged from left to right in an increasing degree of Human-likeness.

Application Context and Acceptance of Vocal Human-Likeness

Regarding context-specific effects, we had hypothesized that, independent from the voice condition, acceptance for the application of a talking robot should be lower for social domains (care, companionship) than for non-social domains (business & finance, information & navigation; *H3a*), whereas with increasing realism and perceived human-likeness of a voice, its acceptance for social applications should increase (*H3b*).

A context-specific mean acceptance index was built by including values of all voice conditions. A Kruskal–Wallis test indicated a significant main effect of application context on user acceptance, $H(5) = 309.599$, $p < 0.001$. In line with *H3a*, this suggests that, independent from the type of voice, application of the talking robot was regarded most acceptable for the less social contexts of "Information & navigation" ($M = 4.07$, SD = 1.14), "Business & finance" ($M = 3.46$, SD = 1.27), "Entertainment" ($M = 3.10$, SD = 1.35), and "Customer service" ($M = 2.84$, SD = 1.30), while study participants had considerably more reservations about its use in the highly social areas "Care" ($M = 1.98$, SD = 1.13) and "Companionship" ($M = 1.68$, SD = 1.06).

Significant differences in user acceptance between the voices could be observed for five out of six contexts (**Figure 3, Table 7 in Appendix C**). A positive correlation between *human-likeness* of the voices and the *context-specific acceptance* was found within all application contexts. The more human-like a voice was perceived, the higher was the acceptance to use the talking service robot in the respective application area. All correlations including a 95% confidence interval based on 1,000 bootstrap samples (Davison and Hinkley, 1997; Shao and Tu, 1995) lie in a range between $r_s = 0.223$, [0.07, 0.38], in the context of "Information & navigation" to $r_s = 0.386$, [0.24, 0.53], in the context of "Care." Having found a positive correlation between human-likeness and user acceptance not specifically within the social domains "Care" and "Companionship" but across all application domains, we regard *H3b* as only partially supported.

User Personality and Acceptance of Vocal Human-Likeness

Finally, we had assumed that individual differences in tolerance of ambiguity and neuroticism would change the nature of the relationship between the perceived human-likeness and eeriness of a voice (H4a, H4b) and that differences in the

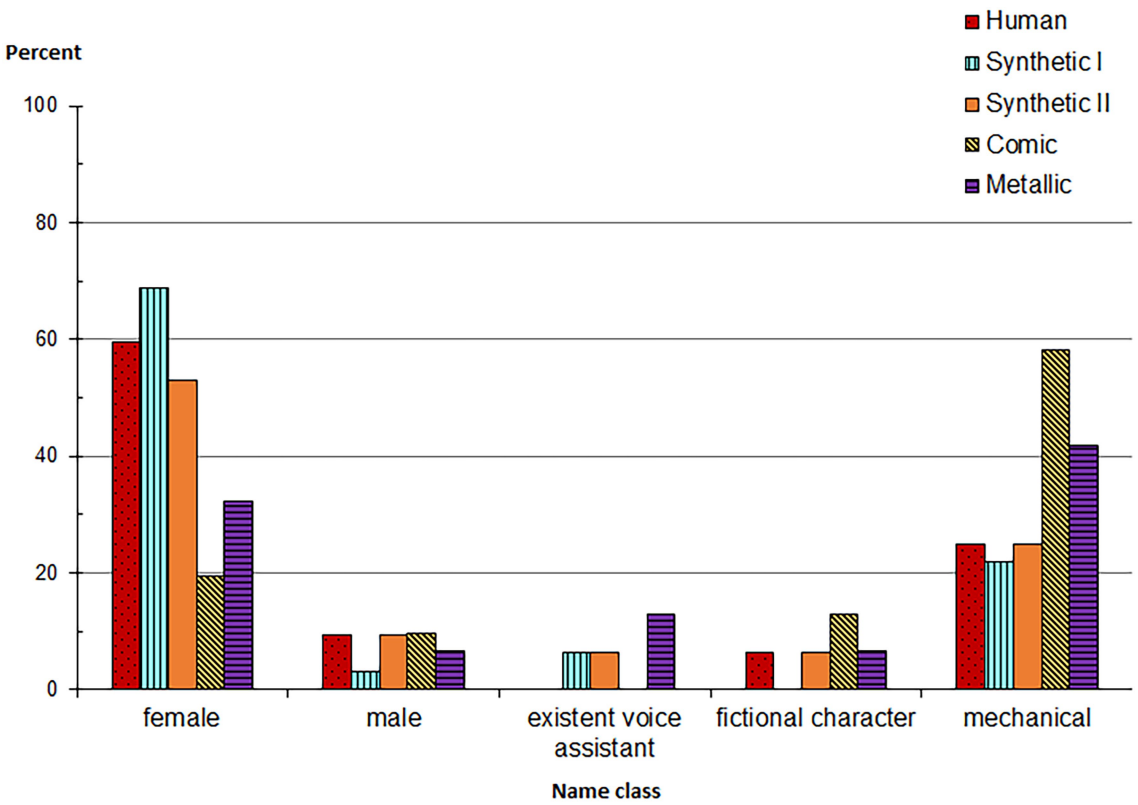

FIGURE 2 | The bar chart shows the absolute values as percentage of invented names depending on the heard voice. The names were assigned to one of the five name classes.

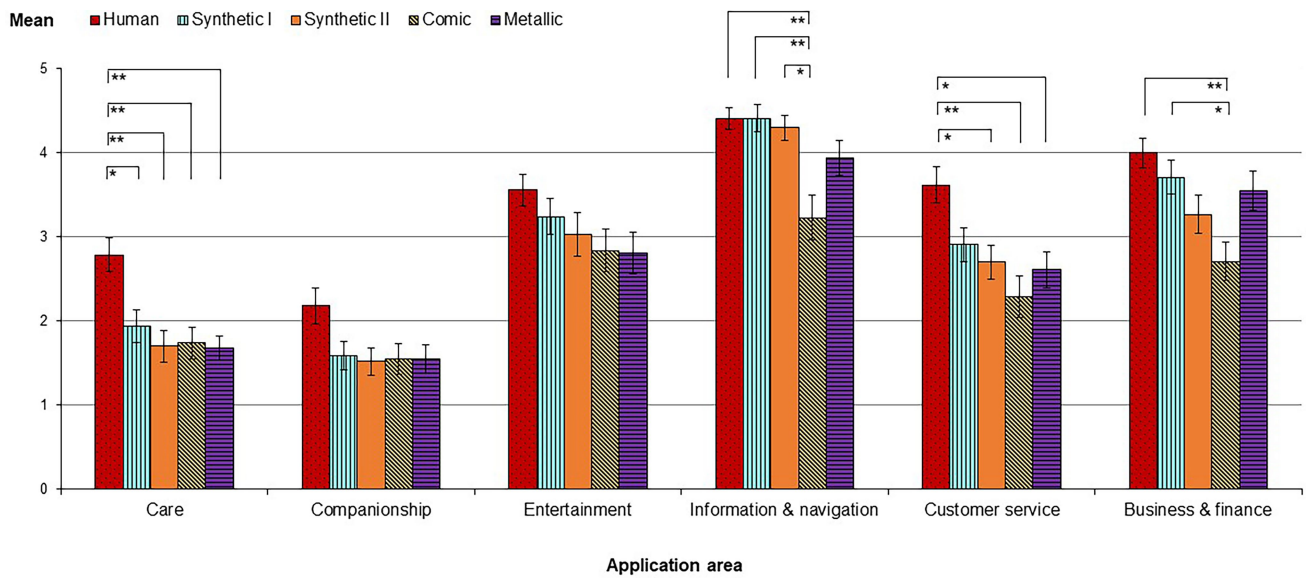

FIGURE 3 | The bar chart shows the mean values of *acceptance* of the five different voices depending on the respective context. A Kruskal–Wallis test was used for pairwise group comparisons (**$p < 0.01$; *$p < 0.05$).

participants' openness to experience would impact the relationship between perceived human-likeness and acceptance of a voice (H4c).

Using the PROCESS macro (version 3.3) for SPSS by Andrew Hayes (Hayes and Cai, 2007; Baltes-Götz, 2017; Process, 2019), we conducted moderation analyses to examine whether *tolerance*

of ambiguity, neuroticism and, for exploratory purposes, the other personality variables of the *Big Five* had a significant influence on the associations between *human-likeness* and *eeriness*. No such interactions on a significance level of $\alpha = 0.05$ were revealed (find more information on the moderation models 1–6 in **Table 8; see Appendix C**). Thus, our hypotheses *H4a* and *H4b* did not find support within this study.

Additionally, moderation models were calculated for the *acceptance* over all contexts (cross-context acceptance index) with *openness to experience* and, for exploratory purposes, the other personality variables as potential moderators. Since the human voice differed significantly from the computer-generated voices in its acceptance, we created a dummy variable (real Human voice vs. all other voices) for model calculation. A confidence level of 95% was set and 5,000 samples were used for bootstrapping. A heteroscedasticity consistent standard error and covariance matrix estimator was used and continuous variables were mean-centered prior to analysis.

In support of *H4c*, a moderation model with robot voice as the predictor (Human vs. all others), *openness to experience* as the moderator, and cross-context acceptance as the outcome variable was found to be significant, $F(3, 159) = 9.63$, $p < 0.01$, $R^2 = 0.15$. A marginal significant interaction $b = 0.35$, $t(159) = 2.01$, $p = 0.046$, indicates a positive influence through higher scores in *openness to experience* on the acceptance of the voice *Human*, but no such effect for the less realistic voices. No moderation effects were found for *tolerance of ambiguity* or the other Big Five dimensions on a significance level of $\alpha = 0.05$ (find more information on moderation models 7–12 in **Table 8; see Appendix C**).

Finally, to check whether participants rated vocal human-likeness differently due to different levels of prior experience, a Kruskal–Wallis test was used to measure the influence of current usage of voice assistant systems ("Are you currently using a voice assistance system at home?") on robot voice acceptance (cross-context). No significant differences were found between those people who are using a voice assistance system, such as Alexa or Siri, and those people who are not (**Table 9; see Appendix C**).

DISCUSSION

The human voice is an essential component of interpersonal communication and a significant influence on the formation of attitudes and opinions about others (Sporer and Schwandt, 2006; Imhof, 2010). In the age of artificial intelligence, attempts are being made to mimic natural language and human voice as closely as possible through technology. Unlike synthetic speech from earlier years, which often failed to produce convincing quality (e.g., Mayer et al., 2003; Atkinson et al., 2005), contemporary computer voices sound more and more natural (Craig and Schroeder, 2017). They prompt the idea that a phone call from a bot, for example, could soon be hardly distinguishable from a real person (Oord et al., 2016; Seaborn and Urakami, 2021)—unless a different design decision is made by the creators of the voice.

User needs and differential preferences should be taken into account early on in technology design. In light of the empirical and theoretical literature presented, however, it was left unclear whether highly realistic sounding synthetic voices were more likely to be linked to positive or negative user responses. With this study, we contribute to the understanding of how different types of voices, supposedly belonging to a service robot, are anthropomorphized, evaluated as pleasant or eerie, and accepted for real-world use. Complementing existing evidence, our randomized experiment for the first time compared assessments of five synthetic voices that differed in their degree of realism while also considering potential influences of contextual (application domain) and dispositional (personality traits) factors.

General Discussion

Consistent with the notion that synthetic voices can serve as major anthropomorphic cues and in support of our Hypotheses 1a–c, more realistic voices were more strongly anthropomorphized than less realistic sounding voices in our experiment. This was expressed not only by higher subjective human-likeness ratings but also by the fact that more realistic voices were more often given a real human name and that study participants also imagined the robot's embodiment to look more human-like. These results are in line with earlier work that revealed object naming as a manifestation of anthropomorphism (Qiu and Benbasat, 2009; Waytz et al., 2010; Brédart, 2021) and they also point us to potential unconscious connections between associative components of auditory and visual stimuli. Further investigations into such associative linkages may be crucial in order to create artificial voices and external object appearances that match each other (Mara et al., 2020). This is underlined by previous research, in which congruent designs of conversational machines were found to contribute to effective interaction and trust (Kiesler and Goetz, 2002; Gong and Nass, 2007; Elkins and Derrick, 2013; Torre et al., 2015, 2018).

Our non-directional Hypotheses 2a–b, stating that there would be significant group differences in pleasantness and eeriness ratings between the voices, found support in such a way that more human-like voices were experienced as significantly more pleasant and less eerie than more mechanical sounding voices. This is in agreement with prior empirical studies that also observed positive effects of anthropomorphic design features (Romportl, 2014; Baird et al., 2018; Kühne et al., 2020; Roesler et al., 2021). At the same time, it seems to contradict the Uncanny Valley hypothesis (Mori, 1970) according to which we would have expected either the quite realistic yet not perfect voices Synthetic I or II receiving the highest eeriness ratings or alternatively—assuming categorical conflicts as an important mechanism behind uncanny experiences—the real human voice (given that participants were told they were listening to a robot). What needs to be noted here is that according to Mori's popular Uncanny Valley graph, which illustrates the assumed curvilinear relationship between human-likeness of a figure and the valence of observer evaluations, a positively valenced peak (most likable, pleasant) should occur at about 70% and a negatively valenced "valley" (most eerie, uncanny) at about 85% along the human-likeness continuum. However,

with a mean value of 3.8 (on a range of 1–5) in reported human-likeness perceptions, even the real human voice in our experiment was relatively far from the right end point of the human-likeness continuum, but closer to the predicted positive peak. From this perspective, by following Mori's postulations, it is not surprising that linear rather than curvilinear relationships between perceived human-likeness and eeriness (or pleasantness) were identified from our data, since the Uncanny Valley hypothesis itself predicts a rather linear increase of positive valence in a low to medium-high range of human similarity, that is, left of the positive peak.

Based on the collected data, it is difficult to answer why the real human voice was not rated as clearly more human-like. Perhaps filtering out the breath sounds in the actor's speech recording (see section "Voice Stimuli") removed an essential feature of human speech, perhaps study participants tried to resolve cognitive dissonance induced by the bad fit of the voice to the label "robot" by reporting lower perceived human-likeness (cf. Festinger, 1962; Marikyan et al., 2020), or perhaps it had to do with the general tendency of study participants to avoid endpoints of response scales (cf. Douven, 2018). A recent meta-analysis on Uncanny Valley effects of embodied humanoid robots suggests that this is a limitation not only of the current work but of many studies in the growing body of related literature. So far, there seem to be hardly any empirical studies that completely cover Mori's human-likeness spectrum or at least make it to the almost-human level with their choice of stimuli (Mara et al., 2022). Future research on Uncanny Valley effects could therefore aim to include stimuli that are closer to the right endpoint of the human-likeness continuum and possibly also pre-test their appropriateness in pilot studies.

Regarding the context-dependent acceptance of robot voices, we found support for our hypothesis H3a. Consistent with previous surveys, in which respondents were significantly more skeptical about the use of robots or AI systems in social applications than in non-social ones (Eurobarometer, 2012; Gaudiello et al., 2016; Ullman et al., 2021), a similar pattern was also reflected in our data. On average across all voices, that is, regardless of their degree of human realism, our participants were significantly more positive about the use of a conversational robot in domains, such as information & navigation or business & finance than in the social–communicative domains care and companionship. In H3b, we had assumed that within these social domains, more human-like voices would yield particularly high acceptance scores due to a perceived congruence between the nature of such voices and typically required "human" skills in this field. After a positive correlation between human-likeness and user acceptance was found not just within social domains but across all included application scenarios, this hypothesis was only partially supported. It is worth noting, however, that the largest correlation coefficient was nonetheless observed in the highly social context of caregiving. However, we cannot completely rule out that the more realistic voices might have been perceived as particularly appropriate for use in social domains, because they also sounded more female than the mechanical voices. Due to prevailing gender stereotypes in society, women are still more often associated with communal traits (e.g., friendly, caring, and gentle) than men (Eagly and Wood, 1982; Hentschel et al., 2019). If voices that sounded more like a real woman were also unconsciously attributed more communal traits in our study, this may have led to a systematic bias in context-specific acceptance scores. To be able to detect such effects, future research is encouraged to include also male-sounding or even gender-neutral synthetic voices (cf. Carpenter, 2019) as stimuli.

While the positive influence of a participant's openness for experience on the acceptance of vocal realism was found in line with H4c, the expected moderating roles of tolerance for ambiguity (H4a) or neuroticism (H4b) in the relationship between human-likeness and perceived eeriness of a voice were not supported by our data. We should note here that both of the latter hypotheses were based on previous findings from the empirical Uncanny Valley literature (MacDorman and Entezari, 2015; Lischetzke et al., 2017), which suggested that individuals with lower tolerance for ambiguity or higher levels of neuroticism would be particularly susceptible to uncanny effects of highly human-like machines. However, with a maximum eeriness rating of 3.45 for the voice Metallic (on a 5-point scale) and much lower eeriness scores for the more realistically sounding voices, no Uncanny Valley effect could be revealed in our study, thus the foundation for the predicted interaction effects was lacking. For individuals with low ambiguity tolerance, our initial assumption was that a possibly perceived conflict between high vocal human-likeness and the simultaneous indication that the speaker is a robot might lead to more pronounced eeriness. Our experimental manipulation did not seem to induce such a conceptual conflict, however. This could be due to the fact that even the real human voice was not rated as very much human-like on average. What, conversely, could have played a role is that a few participants in the Human voice condition expressed disbelief at the end of the study that the voice they had listened could be a robot. Future studies should therefore try to generate more convincing conflicting cues or include a measure for doubt about the presented stimulus as a control variable.

Limitations and Outlook

Beyond the topics discussed above, we note several further limitations of the current study that may at the same time provide suggestions for future research.

First, we were only able to include five stimulus voices in our experiment, which of course cannot cover the full range of existing text-to-speech systems on the market. Although no prior study has compared such a large number of different synthetic voice types, our selection still failed to cover the human-likeness spectrum of Mori's Uncanny Valley graph (Mori, 1970) in the higher third. Hence, it might make sense to elaborate on even more realistic sounding stimuli or on finer gradations along the vocal realism continuum. Instead of features like voice pitch as used in the current study, attempts could be made to manipulate the human-likeness of a talking robot *via* other factors, such as affective content or vocal expression.

Second, we assessed participants' acceptance for the use of the robot voice they had listened to only by means of a self-report scale, which included one item for each application scenario. Although the items were presented in random order within our study, this makes it possible that a participant's different contextual acceptance ratings were not independent of each other. In order to focus more closely on context-specific effects and to investigate them by means of a more rigorous study design, we propose to experimentally manipulate the supposed application area of talking machines in future work. In the frame of the current experiment, given five different voices and six application contexts (5×6 factorial design), this would have required a too large sample size for our lab experiment to ensure sufficient statistical power. However, future studies could focus on a smaller number of voices and create stimulus texts that target different applications for each voice.

Third, we think that the methodological approach of using pre-recorded audio files as experimental stimuli deserves some attention. While we still consider them a straightforward method to keep constant all potential influences (e.g., text content and length) apart from the voice manipulation, unidirectional listening does not represent the typical use case of synthetic voices anymore. To account for the interactivity of today's speech interfaces, it might be worth considering having participants engage in dialog with various synthetic voices or even in live interaction with embodied talking robots.

Fourth, to advance the current line of research, it would also be valuable to go beyond cross-sectional measurements and look at user evaluations over time. Especially with very lifelike synthetic voices, it seems possible that they will raise particularly high expectations about the vividness of human–machine dialogs and the natural language capabilities of the machine. How acceptable or appropriate a synthetic voice is evaluated over time might thus also depend on how much it has been able to withstand such expectations.

Fifth, all participants in our experiment were prepared that they were about to hear a speech recording of a robot. It was not our goal to create ambiguity about the nature of the speaker. This approach is in line with current ethics guidelines for trustworthy technology (High-Level Expert Group on Artificial Intelligence, 2019), which include the requirement that conversational agents should not represent themselves as human but must disclose themselves as machines when communicating with a person. Since it can be assumed that these guidelines will not always be followed in practice, it would be interesting from both a scientific and an applied perspective to see whether a subsequent disclosure—that is, a late notice that a lifelike voice you just listened to was in fact a robot speaking—would trigger more negative user reactions, such as reactance, feelings of a loss of control or uncanny experiences. Thus, even if the participants in this study were relatively welcoming of highly human-like synthetic voices, ethical considerations and psychological consequences of intransparency may still require talking machines to be designed in a way that humans can clearly identify them as such.

CONCLUSION

While technology companies deploy synthetic voices that are barely distinguishable from humans, research on user responses to different grades of vocal human-likeness in machines is still sparse. By testing effects of varying degrees of realism between five robot voices, our findings indicate that robots with more realistic sounding voices are anthropomorphized more strongly, are rated as more pleasant and less eerie, and face the highest acceptance scores across various practical application scenarios. Individuals with high openness for experience were particularly positive about the most human-like voice. Irrespective of the voice type, participants were generally more skeptical of applying talking robots to social domains that, like caregiving, require typically human skills. While this study overall suggests favorable user responses to highly human-like robot voices, a human-centered design of conversational machines certainly requires further research to build on. Beyond our cross-sectional considerations, it remains unclear whether speech interfaces can meet the high user expectations, which are likely to result from lifelike synthetic voices, in the long term. Multidisciplinary research is encouraged to look beyond technical possibilities and psychological effects also at ethical issues, which human-sounding synthetic voices ultimately raise due to their deceptive capacity.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SS and MM contributed to conception and design of the study. SS organized the database and performed the statistical analysis and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank the professional singer Raffaela Gmeiner, whose voice we used for the recordings of the "Human" voice condition. We are also grateful to Benedikt Leichtmann and Christine Busch for their valuable comments on an earlier version of the manuscript and to Franz Berger, who helped

us creating the synthetic voice samples. We would also like to thank Wolfgang Schreibelmayr, who supported us with the acquisition of participants. The publication of this work was supported by the Open Access Publishing Fund of the Johannes Kepler University Linz.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.787499/full#supplementary-material

REFERENCES

Aaltonen, I., Arvola, A., Heikkilä, P., and Lammi, H. (2017). "Hello pepper, may I tickle you? Children's and adults' responses to an entertainment robot at a shopping mall." in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 53–54.

Adobe Audition (2019). [Computer Software]. Available at: https://www.adobe.com/de/products/audition/free-trial-download.html (Accessed April 20, 2019).

Amazon (2017). Available at: https://aws.amazon.com/de/about-aws/whats-new/2017/05/amazon-polly-introduces-a-new-german-female-voice-vicki/ (Accessed August 09, 2021).

Anthony, L. M., Clarke, M. C., and Anderson, S. J. (2000). Technophobia and personality subtypes in a sample of south African university students. *Comput. Hum. Behav.* 16, 31–44. doi: 10.1016/S0747-5632(99)00050-3

Appel, M., Weber, S., Krause, S., and Mara, M. (2016). "On the eeriness of service robots with emotional capabilities." In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 411–412.

Atkinson, R. K., Mayer, R. E., and Merrill, M. M. (2005). Fostering social agency in multimedia learning: examining the impact of an animated agent's voice. *Contemp. Educ. Psychol.* 30, 117–139. doi: 10.1016/j.cedpsych.2004.07.001

Audacity (2019). Available at: https://www.audacityteam.org/ (Accessed April 20, 2019).

Audiveris (2019). Tool for audio-loudness. Available at: https://developers.google.com/actions/tools/audio-loudness (Accessed February 12, 2019).

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., and Schuller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. *Proc. Interspeech* 2018, 2863–2867. doi: 10.21437/Interspeech.2018-1093

Baltes-Götz, B. (2017). *Mediator-und Moderatoranalyse mit SPSS und PROCESS*. Universität Trier Zentrum für Informations-, Medien-und Kommunikationstechnologie (ZIMK).

Bartneck, C., Kanda, T., Ishiguro, H., and Hagita, N. (2007). "Is the uncanny valley an uncanny cliff?" in *RO-MAN 2007-The 16th IEEE international symposium on robot and human interactive communication* (IEEE), 368–373.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi: 10.1007/s12369-008-0001-3

Bendel, O. (2022). *Soziale Roboter: Technikwissenschaftliche, Wirtschaftswissenschaftliche, Philosophische, Psychologische und Soziologische Grundlagen.* Wiesbaden, Germany: Springer Fachmedien Wiesbaden GmbH. doi: 10.1007/978-3-658-31114-8

Blut, M., Wang, C., Wünderlich, N. V., and Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *J. Acad. Mark. Sci.* 49, 632–658. doi: 10.1007/s11747-020-00762-y

Bochner, S. (1965). Defining intolerance of ambiguity. *Psychol. Rec.* 15, 393–400. doi: 10.1007/BF03393605

Brédart, S. (2021). The influence of anthropomorphism on giving personal names to objects. *Adv. Cogn. Psychol.* 17, 33–37. doi: 10.5709/acp-0314-1

Broadbent, E., Jayawardena, C., Kerse, N., Stafford, R. Q., and MacDonald, B. A. (2011). "Human-robot interaction research to improve quality of life in elder care: an approach and issues." in *25th Conference on Artificial Intelligence*. AAAI Workshop, San Francisco, CA.

Burleigh, T. J., Schoenherr, J. R., and Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Comput. Hum. Behav.* 29, 759–771. doi: 10.1016/j.chb.2012.11.021

Carpenter, J. (2019). Why project Q is more than the world's first nonbinary voice for technology. *Interactions* 26, 56–59. doi: 10.1145/3358912

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). "The robotic social attributes scale (RoSAS) development and validation." in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 254–262.

Chang, M., Kim, T. W., Beom, J., Won, S., and Jeon, D. (2020). AI therapist realizing expert verbal cues for effective robot-assisted gait training. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 2805–2815. doi: 10.1109/TNSRE.2020.3038175

Charness, N., Yoon, J. S., Souders, D., Stothart, C., and Yehnert, C. (2018). Predictors of attitudes toward autonomous vehicles: the roles of age, gender, prior knowledge, and personality. *Front. Psychol.* 9:2589. doi: 10.3389/fpsyg.2018.02589

Cohen, J. (1992a). Statistical power analysis. *Curr. Dir. Psychol. Sci.* 1, 98–101. doi: 10.1111/1467-8721.ep10768783

Cohen, J. (1992b). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

Cohn, M., and Zellou, G. (2020). "Perception of concatenative vs. neural text-to-speech (TTS): differences in intelligibility in noise and language attitudes." in INTERSPEECH. 1733–1737.

Costa, P. T., and McCrae, R. R. (1985). The NEO personality inventory.

Couper, M. P., Tourangeau, R., and Steiger, D. M. (2001). "Social presence in web surveys." in *Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference*. eds. M. Beaudouin-Lafon, J. Beaudouin-Lafon, and J. K. Robert (New York: ACM Press), 412–417.

Craig, S. D., and Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Comput. Educ.* 114, 193–205. doi: 10.1016/j.compedu.2017.07.003

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13:319. doi: 10.2307/249008

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and their Application (No. 1)*. United Kingdom: Cambridge University Press. doi: 10.1017/CBO9780511802843

De Graaf, M. M., Ben Allouch, S., and Van Dijk, J. A. (2015). "What makes robots social?: a user's perspective on characteristics for social human-robot interaction." in *International Conference on Social Robotics* (Cham: Springer), 184–193.

Devaraj, S., Easley, R. F., and Crant, J. M. (2008). Research note—how does personality matter? Relating the five-factor model to technology acceptance and use. *Inf. Syst. Res.* 19, 93–105. doi: 10.1287/isre.1070.0153

Diel, A., and MacDorman, K. F. (2021). Creepy cats and strange high houses: support for configural processing in testing predictions of nine uncanny valley theories. *J. Vis.* 21:1. doi: 10.1167/jov.21.4.1

Digman, J. M. (1990). Personality structure: emergence of the five-factor model. *Annu. Rev. Psychol.* 41, 417–440. doi: 10.1146/annurev.ps.41.020190.002221

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Psychon. Bull. Rev.* 25, 1203–1211. doi: 10.3758/s13423-017-1344-2

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robot. Auton. Syst.* 42, 177–190. doi: 10.1016/S0921-8890(02)00374-3

Eagly, A. H., and Wood, W. (1982). Inferred sex differences in status as a determinant of gender stereotypes about social influence. *J. Pers. Soc. Psychol.* 43, 915–928. doi: 10.1037/0022-3514.43.5.915

Elkins, A. C., and Derrick, D. C. (2013). The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group Decis. Negot.* 22, 897–913. doi: 10.1007/s10726-012-9339-x

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* 114, 864–886. doi: 10.1037/0033-295X.114.4.864

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2021). "A meta-analysis of human personality and robot acceptance in human-robot

interaction." in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

Eurobarometer Special 382 (2012). Public attitudes towards robots. European Commission. Available at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_sum_en.pdf (Accessed March 4, 2021).

Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., and Hegel, F. (2012). "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism." in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. (IEEE), 125–126.

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Festinger, L. (1962). Cognitive dissonance. *Sci. Am.* 207, 93–106. doi: 10.1038/scientificamerican1062-93

Fink, J., Mubin, O., Kaplan, F., and Dillenbourg, P. (2012). "Anthropomorphic language in online forums about Roomba, AIBO and the iPad." in *2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (IEEE), 54–59.

Freeston, M. H., Rhéaume, J., Letarte, H., Dugas, M. J., and Ladouceur, R. (1994). Why do people worry? *Personal. Individ. Differ.* 17, 791–802. doi: 10.1016/0191-8869(94)90048-5

Furnham, A., and Ribchester, T. (1995). Tolerance of ambiguity: A review of the concept, its measurement and applications. *Curr. Psychol.* 14, 179–199. doi: 10.1007/BF02686907

Gambino, A., and Sundar, S. S. (2019). "Acceptance of self-driving cars: does their posthuman ability make them more eerie or more desirable?." in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., and Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Comput. Hum. Behav.* 61, 633–655. doi: 10.1016/j.chb.2016.03.057

Giles, H., Scherer, K. R., and Taylor, D. M. (1979). "Speech markers in social interaction," in *Social markers in speech*. eds. K. R. Scherer and H. Giles (Cambridge, UK: Cambridge University Press), 343–381.

Goetz, J., Kiesler, S., and Powers, A. (2003). "Matching robot appearance and behavior to tasks to improve human-robot cooperation." in *12th IEEE International workshop on robot and human interactive communication*, IEEE, Millbrae, CA, 55–60.

Gong, L., and Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: human identity and consistency preference. *Hum. Commun. Res.* 33, 163–193. doi: 10.1111/j.1468-2958.2007.00295.x

Google Duplex (2018). A.I. Assistant Calls Local Businesses To Make Appointments. Available at: https://www.youtube.com/watch?v=D5VN56jQMW (Accessed May 5, 2018).

Hayes, A. F., and Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation. *Behav. Res. Methods* 39, 709–722. doi: 10.3758/BF03192961

Hedda (2019). Microsoft speech platform. Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/quickstart-python-text-to-speech (Accessed March 20, 2019).

Hentschel, T., Heilman, M. E., and Peus, C. V. (2019). The multiple dimensions of gender stereotypes: a current look at men's and women's characterizations of others and themselves. *Front. psychol.* 10:11. doi: 10.3389/fpsyg.2019.00011

High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*.

Ho, C. C., and MacDorman, K. F. (2010). Revisiting the uncanny valley theory: developing and validating an alternative to the Godspeed indices. *Comput. Hum. Behav.* 26, 1508–1518. doi: 10.1016/j.chb.2010.05.015

Ho, C. C., and MacDorman, K. F. (2017). Measuring the uncanny valley effect. *Int. J. Soc. Robot.* 9, 129–139. doi: 10.1007/s12369-016-0380-9

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *J. R. Stat. Soc. Ser. B Methodol.* 30, 582–598. doi: 10.1111/j.2517-6161.1968.tb00759.x

Ilves, M., and Surakka, V. (2013). Subjective responses to synthesised speech with lexical emotional content: the effect of the naturalness of the synthetic voice. *Behav. Inform. Technol.* 32, 117–131. doi: 10.1080/0144929X.2012.702285

Imhof, M. (2010). *Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht*. 15–30.

Jia, J. W., Chung, N., and Hwang, J. (2021). Assessing the hotel service robot interaction on tourists' behaviour: the role of anthropomorphism. *Ind. Manag. Data Syst.* 121, 1457–1478. doi: 10.1108/IMDS-11-2020-0664

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). *The "Big Five" Inventory – Versions 4a and 54. Technical Report*. Berkeley: University of California, Institute of Personality and Social Research.

Jung, Y., and Cho, E. (2018). "Context-specific affective and cognitive responses to humanoid robots." in *The 22nd biennial conference of the international telecommunications society: "beyond the boundaries: Challenges for business, policy and society."* International Telecommunications Society (ITS), Seoul, Korea.

Juniper (2019). [Online research platform]. Available at: https://www.juniperresearch.com/press/press-releases/digital-voice-assistants-in-use-to-8-million-2023 (Accessed December 9, 2019).

Kaplan, P. S., Goldstein, M. H., Huckeby, E. R., and Cooper, R. P. (1995). Habituation, sensitization, and infants' responses to motherse speech. *Dev. Psychobiol.* 28, 45–57. doi: 10.1002/dev.420280105

Kaur, R., Sandhu, R. S., Gera, A., Kaur, T., and Gera, P. (2020). "Intelligent voice bots for digital banking," in *Smart Systems and IoT: Innovations in Computing*. eds. A. K. Somani, R. S. Shekhawat, A. Mundra, S. Srivastava and V. K. Verma (Singapore: Springer), 401–408. doi: 10.1007/978-981-13-8406-6_38

Kiesler, S., and Goetz, J. (2002). "Mental models of robotic assistants." in *CHI'02 extended abstracts on human factors in computing systems*, 576–577.

Kohlberg, L., DeVries, R., Fein, G. G., Hart, D., Mayer, R., Noam, G. G., et al. (1987). *Child Psychology and Childhood Education: A Cognitive-Developmental View*. United States: Addison-Wesley Longman Limited.

Krauss, R. M., Freyberg, R., and Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *J. Exp. Soc. Psychol.* 38, 618–625. doi: 10.1016/S0022-1031(02)00510-3

Kühne, K., Fischer, M. H., and Zhou, Y. (2020). The human takes it All: humanlike synthesized voices are perceived as less eerie and more likable. Evidence From a subjective ratings study. *Front. Neurorobot.* 14:105. doi: 10.3389/fnbot.2020.593732

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lischetzke, T., Izydorczyk, D., Hüller, C., and Appel, M. (2017). The topography of the uncanny valley and individuals' need for structure: a nonlinear mixed effects analysis. *J. Res. Pers.* 68, 96–113. doi: 10.1016/j.jrp.2017.02.001

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., et al. (2019). Talk to me: exploring user interactions with the Amazon Alexa. *J. Librariansh. Inf. Sci.* 51, 984–997. doi: 10.1177/0961000618759414

MacDorman, K. F., and Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interact. Stud.* 16, 141–172. doi: 10.1075/is.16.2.01mac

Mara, M., and Appel, M. (2015a). Effects of lateral head tilt on user perceptions of humanoid and android robots. *Comput. Hum. Behav.* 44, 326–334. doi: 10.1016/j.chb.2014.09.025

Mara, M., and Appel, M. (2015b). Science fiction reduces the eeriness of android robots: a field experiment. *Comput. Hum. Behav.* 48, 156–162. doi: 10.1016/j.chb.2015.01.007

Mara, M., Appel, M., and Gnambs, T. (2022). Human-like robots and the uncanny valley: a meta-analysis of user responses based on the godspeed scales. *Z. Psychol.* 230, 33–46. doi: 10.1027/2151-2604/a000486

Mara, M., Schreibelmayr, S., and Berger, F. (2020). "Hearing a nose? User expectations of robot appearance induced by different robot voices." in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 355–356.

Maricutoiu, L. P. (2014). A meta-analysis on the antecedents and consequences of computer anxiety. *Procedia Soc. Behav. Sci.* 127, 311–315. doi: 10.1016/j.sbspro.2014.03.262

Marikyan, D., Papagiannidis, S., and Alamanos, E. (2020). Cognitive dissonance in technology adoption: a study of smart home users. *Inf. Syst. Front.* 1–23. doi: 10.1007/s10796-020-10042-3

Mathur, M. B., and Reichling, D. B. (2016). Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. *Cognition* 146, 22–32. doi: 10.1016/j.cognition.2015.09.008

Mayer, R. E., Sobko, K., and Mautone, P. D. (2003). Social cues in multimedia learning: role of speaker's voice. *J. Educ. Psychol.* 95, 419–425. doi: 10.1037/0022-0663.95.2.419

McGee, T. J., King, C., Tremblay, K., Nicol, T. G., Cunningham, J., and Kraus, N. (2001). Long-term habituation of the speech-elicited mismatch negativity. *Psychophysiology* 38, 653–658. doi: 10.1111/1469-8986.3840653

Meah, L. F., and Moore, R. K. (2014). "The uncanny valley: a focus on misaligned cues." in *International Conference on Social Robotics* (Cham: Springer), 256–265.

Meinecke, C. (2019). Available at: https://www2.deloitte.com/at/de.html (Accessed October 14, 2019).

Mejia, C., and Kajikawa, Y. (2017). Bibliometric analysis of social robotics research: identifying research trends and knowledgebase. *Appl. Sci.* 7:1316. doi: 10.3390/app7121316

Mitchell, W. J., Szerszen, K. A. Sr., Lu, A. S., Schermerhorn, P. W., Scheutz, M., and MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2, 10–12. doi: 10.1068/i0415

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy* 7, 33–35.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Morsunbul, U. (2019). Human-robot interaction: how do personality traits affect attitudes towards robot? *J. Hum. Sci.* 16, 499–504. doi: 10.14687//jhs.v16i2.5636

Nass, C., and Brave, S. (2005). *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*. Cambridge: MIT press.

Nass, C., and Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *J. Exp. Psychol. Appl.* 7, 171–181. doi: 10.1037/1076-898X.7.3.171

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 72–78.

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., and See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* 5, 171–191. doi: 10.1007/s12369-012-0171-x

Norton, R. W. (1975). Measurement of ambiguity tolerance. *J. Pers. Assess.* 39, 607–619. doi: 10.1207/s15327752jpa3906_11

Nov, O., and Ye, C. (2008). "Personality and technology acceptance: personal innovativeness in IT, openness and resistance to change." in *Proceedings of the 41st annual Hawaii International Conference on System Sciences (HICSS 2008)* (IEEE), 448.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). Wavenet: A generative model for raw audio [Preprint]. arXiv arXiv:1609.03499.

Oshio, A. (2009). Development and validation of the dichotomous thinking inventory. *Soc. Behav. Pers.* 37, 729–741. doi: 10.1037/t68911-000

Oyedele, A., Hong, S., and Minor, M. S. (2007). Contextual factors in the appearance of consumer robots: exploratory assessment of perceived anxiety toward humanlike consumer robots. *CyberPsychol. Behav.* 10, 624–632. doi: 10.1089/cpb.2007.9977

Pérula-Martínez, R., Castro-González, Á., Malfaz, M., and Salichs, M. A. (2017). "Autonomy in human-robot interaction scenarios for entertainment." in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 259–260.

Pinker, S. (2003). *The Language Instinct: How the Mind Creates Language*. UK: Penguin.

Polly (2019). Polly voices. Available at: https://ttsmp3.com/ (Accessed March 20, 2019).

Process (2019). SPSS Macro [software implementation]. Available at: https://processmacro.org/index.html (Accessed May 10, 2019).

Qiu, L., and Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: a social relationship perspective to designing information systems. *J. Manag. Inf. Syst.* 25, 145–182. doi: 10.2753/MIS0742-1222250405

Questback (2018). Available at: https://www.questback.com/de/ (Accessed December 11, 2018).

Radant, M., and Dalbert, C. (2003). Zur Dimensionalität der Ambiguitätstoleranz. Poster auf der 7. DPPD-Tagung der Deutschen Gesellschaft für Psychologie.

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People*. Cambridge, UK: CSLI Publications.

Robinson, M. D. (2004). Personality as performance: categorization tendencies and their correlates. *Curr. Dir. Psychol. Sci.* 13, 127–129. doi: 10.1111/j.0963-7214.2004.00290.x

Robinson, M. D., Vargas, P. T., and Crawford, E. G. (2003). "Putting process into personality, appraisal, and emotion: evaluative processing as a missing link," in *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*. eds. J. Musch and K. C. Klauer (Mahwah, NJ: Lawrence Erlbaum), 275–306. doi: 10.4324/9781410606853-19

Roesler, E., Manzey, D., and Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Sci. Robot.* 6:eabj5425. doi: 10.1126/scirobotics.abj5425

Roesler, E., Naendrup-Poell, L., Manzey, D., and Onnasch, L. (2022). Why context matters: the influence of application domain on preferred degree of anthropomorphism and gender attribution in human–robot interaction. *Int. J. Soc. Robot.* 14, 1–12. doi: 10.1007/s12369-021-00860-z

Romportl, J. (2014). "Speech synthesis and uncanny valley" in *International Conference on Text, Speech, and Dialogue* (Cham: Springer), 595–602.

Schlink, S., and Walther, E. (2007). Kurz und gut: Eine deutsche Kurzskala zur Erfassung des Bedürfnisses nach kognitiver Geschlossenheit. *Z. Sozialpsychol.* 38, 153–161. doi: 10.1024/0044-3514.38.3.153

Schupp, J., and Gerlitz, J. Y. (2014). "Big Five Inventory-SOEP (BFI-S)." in *Zusammenstellung sozialwissenschaftlicher Items und Skalen* (Vol. 10).

Seaborn, K., and Urakami, J. (2021). "Measuring voice UX quantitatively: a rapid review." in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

Shao, J., and Tu, D. (eds.) (1995). "Theory for the Jackknife," in *The Jackknife and Bootstrap. Springer Series in Statistics* (New York: Springer), 23–70. doi: 10.1007/978-1-4612-0795-5_2

Smith, H. M., Dunn, A. K., Baguley, T., and Stacey, P. C. (2016). Concordant cues in faces and voices: testing the backup signal hypothesis. *Evol. Psychol.* 14:147470491663031. doi: 10.1177/1474704916630317

Sporer, S. L., and Schwandt, B. (2006). Paraverbal indicators of deception: a meta-analytic synthesis. *Appl. Cogn. Psychol.* 20, 421–446. doi: 10.1002/acp.1190

Sprent, P. (2007). An introduction to categorical data analysis. *J. R. Stat. Soc. Ser. A* 170:1178. doi: 10.1111/j.1467-985X.2007.00506_2.x

Statista (2021). [Online research platform]. Available at: https://www.statista.com (Accessed March 10, 2021).

Sutton, S. J., Foulkes, P., Kirk, D., and Lawson, S. (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*. 1–14.

Tiwari, M., and Tiwari, M. (2012). Voice-how humans communicate? *J. Nat. Sci. Biol. Med.* 3, 3–11.

Torre, I., Goslin, J., and White, L. (2015). "Investing in accents: how does experience mediate trust attributions to different voices?" in *ICPhS*.

Torre, I., Goslin, J., White, L., and Zanatto, D. (2018). "Trust in artificial voices: a "congruency effect" of first impressions and behavioural experience." in *Proceedings of the Technology, Mind, and Society*. 1–6.

Tourangeau, R., Couper, M. P., and Steiger, D. M. (2003). Humanizing self-administered surveys: experiments on social presence in web and IVR surveys. *Comput. Hum. Behav.* 19, 1–24. doi: 10.1016/S0747-5632(02)00032-8

Ullman, D., Aladia, S., and Malle, B. F. (2021). "Challenges and opportunities for replication science in HRI: a case study in human-robot trust." in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 110–118.

Vlachos, E., Jochum, E., and Demers, L. P. (2016). The effects of exposure to different social robots on attitudes toward preferences. *Interact. Stud.* 17, 390–404. doi: 10.1075/is.17.3.04vla

Voxal (2019). Voice changer. Available at: https://www.nchsoftware.com/voicechanger/de/index.htm (Accessed January 15, 2019).

Wada, K., and Shibata, T. (2006). "Living with seal robots in a care house-evaluations of social and physiological influences." in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE), 4940–4945.

Wada, K., Shibata, T., Saito, T., and Tanie, K. (2003). "Effects of robot assisted activity to elderly people who stay at a health service facility for the aged." in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent*

Robots and Systems (IROS 2003)(Cat. No. 03CH37453). Vol. 3 (IEEE), 2847–2852).

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* 5, 219–232. doi: 10.1177/1745691610369336

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education. Technical Report. UNESCO, EQUALS Skills Coalition. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000367416.locale=en (Accessed April 1, 2022).

Whang, C., and Im, H. (2021). "I like your suggestion!" the role of humanlikeness and parasocial relationship on the website versus voice shopper's perception of recommendations. *Psychol. Mark.* 38, 581–595. doi: 10.1002/mar.21437

Zhang, T., Tao, D., Qu, X., Zhang, X., Zeng, J., Zhu, H., et al. (2020). Automated vehicle acceptance in China: social influence and initial trust are key determinants. *Transp. Res. Part C Emerg. Technol.* 112, 220–233. doi: 10.1016/j.trc.2020.01.027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.