Computers in Human Behavior 145 (2023) 107791

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude

Hyunjoo Im [a,*], Billy Sung [b], Garim Lee [a], Keegan Qi Xian Kok [b]

[a] Retailing and Consumer Studies, University of Minnesota – Twin Cities, 240 McNeal Hall, 1985 Buford Ave., St. Paul, MN, 55108, USA
[b] Consumer Research Lab, School of Management and Marketing, Curtin University, Bently, WA, 6102, Australia

ARTICLE INFO

Handling Editor: Marianna Sigala

Keywords:
Voice assistant
Synthetic voice
Voice perception

ABSTRACT

Voice Assistants (VA) are increasingly penetrating consumers' daily lives. This study aimed to investigate the effects of synthetic vs. human voice on users' perception of the voice, social judgments of the VA, and attitudes towards VAs. Drawing from CASA(Computers-as-socialactors) framework and social perception literature, we developed a theoretical model that explains the psychological underlying mechanism of the voice effects. Through two online experiments, we rejected our initial hypotheses that human voice would increase users' perception and evaluation of the VAs. Instead, our findings support that the VAs were favored when they spoke with a synthetic voice only when the users engaged in functional tasks. There was no difference between the voices for social tasks. A further investigation revealed that participants perceived the synthetic voice to be more fluent when VA responds to functional tasks. This enhanced perception of fluency increased competence perception and attitudes. The findings imply that VAs should not be designed to closely resemble humans. Rather, consideration of usage contexts and consumer expectations should be prioritized in developing most likable VAs.

1. Introduction

Voice assistants (VAs) such as Amazon Alexa, Google Assistant, and Apple Siri have become increasingly prevalent in everyday life. VAs are becoming omnipresent as they are embedded as a chief user interface mechanism in smart devices at home (e.g., smart TVs, smart speakers, smartphones), car entertainment systems, or service robots. Approximately 4.2 billion VAs are being used in devices throughout the world in the year 2020 and it is predicted that, by the year 2024, 8.4 billion VAs will be used globally (Number of voice assistants in use worldwide, 2019–2024, n.d.). As a result, VAs are integrated into consumers' daily activities and voice-activated interactions with devices are becoming ever more popular.

Still, interactions with VAs are not as natural as human-to-human interactions. Consumers are reportedly dissatisfied with VAs due to their limited functionality, failed or irrelevant responses, and unnatural conversations (PricewaterhouseCoopers, 2018; Williams, 2019). These unsuccessful interactions make many users doubt the VAs and further negatively impact consumers' trust and expanded use of the VAs beyond the basic tasks (PricewaterhouseCoopers, 2018).

Some identified the synthetic voice as one of the reasons for poor interaction quality. Synthetic voices are perceived as unnatural and unattractive because they lack emotions and intonation (Dou et al., 2020). The robotic, emotionless voice can make the VAs less humanlike and devoid of personality, and is one important reason for chatbot fails (Coredna, 2021). As VAs use voice to verbally deliver the text response to the commands, how the synthetic voice is designed and how consumers perceive the synthetic voice is going to be a critical factor for VA performance.

A substantial amount of research exists in voice design and human-computer interactions on the topic of user experience and perception of synthetic voices. In this stream of research, researchers evaluated the synthetic voice in comparison to the human voice to make recommendations for voice design (Torre & White, 2021). These studies compared synthetic voices with recorded human voices, conceptually manipulating the degree of humanlikeness, naturalness, or realism (Chérif & Lemoine, 2019; Fan et al., 2016; Higgins et al., 2022; Kühne et al., 2020; Schreibelmayr & Mara, 2022). Collectively, these studies support that humanlikeness of voice is a positive factor to induce favorable user responses.

While these studies provide valuable insights for voice design and how users perceive synthetic and human voices, there are still some

* Corresponding author.
E-mail addresses: hjim@umn.edu (H. Im), billy.sung@curtin.edu.au (B. Sung), lee02169@umn.edu (G. Lee), keegan.kok@curtin.edu.au (K.Q. Xian Kok).

https://doi.org/10.1016/j.chb.2023.107791
Received 23 September 2022; Received in revised form 6 April 2023; Accepted 16 April 2023
Available online 18 April 2023
0747-5632/Published by Elsevier Ltd.

knowledge gaps in the literature to fully understand how consumers respond to VA voices. First of all, previous studies generally focused on the comparison of voice perception rather than the psychological underpinning of the voice effects, and therefore failed to investigate the exact underlying mechanisms of the phenomena. Secondly, previous findings may not be directly applicable to the consumer-VA interaction context. Some researchers suggested that voice perception may be fundamentally different depending on whether the voice assistant is embodied (e.g., humanoid service robots speaking with a synthetic vs. human voice) or disembodied (e.g., Siri on iPhone) (Cambre & Kulkarni, 2019). As voice perception was of particular interest to social robot designers, the majority of research in this field was conducted in the context of human-robot interaction, and the voice was often clearly embodied. In contrast, the current study is interested in disembodied agents. Thirdly, despite of the influential studies that demonstrated the importance of usage context (Torre et al., 2020), it was often neglected in the previous studies. When usage contexts were investigated along with the voice, they were selected to test the performance of robots in most likely applications (e.g., hotel reception, hospital, school) (Dou et al., 2020; Torre et al., 2020). These usage contexts are different from more private and personal usage situations of VAs and the findings may not be applicable to consumer evaluation of VAs. Moreover, because the contexts were selected for practical rather than theoretical purposes, the comparisons among the contexts cannot be understood systematically.

Therefore, there is a need to investigate how consumers perceive and respond to VAs with different levels of humanlikeness and explain the perception and evaluation process systematically. This study relies heavily on the social cognition literature while synthesizing voice design research and human-robot interaction research to propose a social psychological explanation of consumer responses to VAs. Building on the robust evidence of the previous research supporting the positive effects of humanlikeness on consumer responses (e.g., Chérif & Lemoine, 2019; Schreibelmayer & Mara, 2022), the current study aims to investigate which humanlike voice quality perception (e.g., voice naturalness) drives the effects. Specifically, the purpose of the study is to investigate 1) to what extent consumers use the voice (i.e., synthetic vs. natural human voice) to make social judgments of the VAs, 2) which perception of the voice characteristics explains the difference the synthetic vs. human voice produces, and 3) whether the usage contexts (i.e., social vs. functional) amplifies or diminishes the effects of voice type.

2. Literature review and hypotheses

2.1. Human vs. synthetic voice

As VAs predominantly, if not entirely, rely on voice to interact with the users, the voice can have a significant impact on users' perception of and attitude towards VAs. VAs use a synthesized voice which is artificially generated by converting text to speech using the technique of selecting and joining words and sounds. The sentences are created by joining pre-recorded words and diphones stored in a database. Although recent technological development improved the quality of synthesized voices to be closer to natural human speech than before (McDonough, 2020), evidence to date shows synthesized voices still sound unnatural or robotic (Rella, 2021) because of the differences or the lack of paralinguistic voice features such as breathing, intonation, and emotional expression that determine the perceived quality of the synthesized voice. As a result, people not only distinguish human voice from synthesized voice but also prefer human voice to synthesized voice (Kühne et al., 2020).

Several empirical studies confirmed this superior effect of human voice. In studies that investigated the impact of humanlikeness of the voice of virtual agents on people's perception and social judgment consistently reported that users responded less favorably to a virtual agent with a synthetic voice than an agent with a human voice (Chérif & Lemoine, 2019; Craig et al., 2019; Stern et al., 2006). People also found a human voice more pleasant (Stern et al., 2006), likable (Kühne et al., 2020), and trustworthy (Chérif & Lemoine, 2019; Craig et al., 2019; Kühne et al., 2020) than a synthetic voice. While some researchers (Kühne et al., 2020; Schreibelmayr & Mara, 2022) considered the possible negative effect when the voice of the non-human agents is too humanlike (based on the Uncanny Valley hypothesis), the results did not support the negative effect of extreme humanlikeness in the context of voice (Schreibelmayr & Mara, 2022). A study that investigated 12 synthetic voices with varying degrees of humanlikeness and 1 human voice also confirmed that humanlikeness of the voice is linearly and positively correlated with likeability of the voice (Baird et al., 2018). Thus, the literature so far collectively suggests that VAs with a human voice (vs. synthetic voice) would be more favorably evaluated by consumers. Therefore, H1 was formulated.

H1. Consumers interacting with a VA with a human (vs. synthetic) voice will report a more (less) positive attitude towards VA.

2.2. Underlying mechanism of the human voice effect

While the literature to date suggests that people will like the VAs more when the voice of the VAs is very humanlike, the literature seldom clarified how humanlikeness positively influences the users (i.e., internal responses that lead to favorable outcomes). In this study, we argue that users of VAs, due to the nature of conversational interactions between the user and the VAs, automatically use the human-to-human social interaction perception and judgments and that the human (vs. synthetic) voice of VAs can create more natural user-VA interactions. As a result, the human voice is inherently more likely to trigger emotional responses that natural interpersonal social interactions typically elicit, which in turn produces a favorable response to the VAs than the synthesized voice does. In the following section, we will first discuss the theoretical framework, CASA (Computers-As-Social-Actors), that explains why VA users apply social perception and judgment rules to VAs. Then, we will explain the psychological outcomes of treating VAs as social actors.

2.2.1. CASA (Computers-As-Social-Actors) framework and voice as the social cue

Humans treat people and objects (e.g., people in TV, puppets, computers) that display social behaviors as if they are actual humans because human brains do not distinguish actual people from mediated representations of people (Reeves & Nass, 1996). The CASA framework confirms this 'media equation' behaviors of humans in the context of human-computer interactions and proposes that humans subconsciously and automatically apply social rules to computers and respond to computers as interaction partners (Gambino et al., 2020; Nass & Moon, 2000). The CASA framework has guided many studies to explain how users behave and interact with computers (Nass & Lee, 2001; Nijholt, 2003; Wang, 2017). Existent literature provides robust evidence to support the main premise of the CASA framework. Participants in previous studies automatically applied social rules they would typically use in human-to-human social interactions when interacting with computers (Lee & Nass, 2010; Nass & Lee, 2001). A few recent studies reported corroborating results for VAs, providing supporting evidence that people treat VAs as social interaction partners (Carolus et al., 2021; Seymour & Van Kleek, 2021; Whang & Im, 2021). The current study shares the view of the CASA framework and assumes that consumers interact and respond to the VAs as if they interact with another person.

The presence of social cues, even the minimal cues, encourages and elicits social responses from humans (Voorveld & Araujo, 2020). In the context of VAs, the voice itself can be a social cue. Humans constantly extract social information from voice of others (Latinus & Belin, 2011) and automatically and subconsciously use voices to infer various human characteristics such as age, gender, personality, and identity (Imhof, 2010). People extend this social perception process to machines and use

the voice of the computers similarly to form impressions and to determine how to react to the machines (Nass & Moon, 2000). For example, people infer personality traits from the synthesized computer voice, form impressions of "the speaker", and determine how trustworthy or likable the speaker (i.e., the computer) is (Nass & Lee, 2001). Likewise, recent research on user perception of robots showed that the design of voice can affect people's perception of and intention toward robots (Dou et al., 2020; Niculescu et al., 2013). When interacting with AI agents, humans use the voice as a source of information to assess social qualities of the AI (Krenn et al., 2017).

Evidence from human-computer interaction and consumer research demonstrates that humans treat non-human objects and machines like human counterparts (Mourey et al., 2017; Nass & Moon, 2000). As a result, people often perceive and respond to non-human machines such as VAs in much the same way as they do to other humans (Whang & Im, 2021). Therefore, it is possible that consumer-VA interactions can be understood through the principles of social cognition (Pitardi & Marriott, 2021).

2.2.2. Social perception: warmth and competence

People can assess the other partner in a social situation almost instantaneously using various cues such as appearance, facial expression, voice, gesture, etc. In the seminal work, Fiske et al. (2007) concluded that warmth and competence are two fundamental dimensions of person perception that further determine how a person responds to the other in an interaction. Warmth is a measure of personality attributes such as friendliness, helpfulness, and sincerity, whereas competence is a dimension of a person's skill, ability, and intelligence. Previous studies also showed that warmth and competence perceptions are key indicators to predict individuals' responses to non-human machines, including the perceived value of, positive attitude towards, and intention to keep interacting with the machine (Belanche et al., 2021; Fernandes & Oliveira, 2021). Extending these findings to VAs, consumers are likely to quickly judge warmth and competence of VAs from the voice of VAs and consequently form attitudes towards the VA.

In this vein, if consumers form more positive attitudes towards the VA that uses a human (vs. synthetic) voice (H1), the preference for the human voice is likely to be explained by heightened perceptions of warmth and competence of the VA with a human voice. While people treat machines as a social actor and apply social judgment rules, research also demonstrated that meaningful differences in warmth and competence perception exist between when they interact with real humans versus computers. For example, Chérif and Lemoine (2017) confirmed that people perceive a higher warmth (conceptualized as warm social presence) when listening to a human voice than a synthetic voice. Such heightened warm social presence is likely to promote individuals to feel human traits such as warmth and sociability more strongly. Also, it has been consistently confirmed that individuals do not perceive machinelike robots as competent as humanlike robots (Stern et al., 2006; Stroessner & Benitez, 2019). Instead, several studies on robot perceptions demonstrated that increased humanlikeness in robots' appearance and voices enhances the perception of warmth and competence of the robots (Kim et al., 2019; Pitardi & Marriott, 2021; Vernuccio et al., 2022). Taken together, consumers are likely to perceive a VA with a human voice as warmer and more competent than a VA with a synthetic voice, which further enhances their attitude toward the VA.

H2. Warmth and competence will mediate the positive human voice effect on attitude.

2.2.3. Perceived naturalness of voice

Natural-sounding voice was at the heart of synthetic voice development for a long time because perceived naturalness of voice has been identified as a critical determinant of many important outcomes such as acceptability of the voice (Nusbaum et al., 1995). Consequently, increasing naturalness of the voice interaction was one of the ultimate goals of the voice user interface design (Fraj et al., 2009; Urakami et al., 2020). While what makes voice natural is very complex and is beyond the scope of the current study, research so far agreed that perceived naturalness is a valid way to capture the overall quality of synthetic voice and that people associate perceived naturalness of voice with humanlikeness of voice (Kühne et al., 2020; McGinn & Torre, 2019). Studies which compared the human voice with the synthetic voice confirmed that people perceive a human actor's voice as more natural than a synthetically produced voice (e.g., Kühne et al., 2020). Thus, a synthetic voice of VAs will likely sound unnatural compared to a human voice. When the voice of a VA is unnatural, users quickly recognize it and may not be able to intuitively and smoothly engage in conversations with the VA as they do with other human beings. Thus, we propose perceived naturalness of voice as a key predictor of the social cognition process (i.e., judgments of warmth and competence) in the user-VA interactions.

Importantly, perceived naturalness of VAs' voices is likely to enhance attitude towards VAs by increasing positive social perceptions of VAs rather than directly influencing attitudes toward VAs. In other words, the superior human voice effect on attitude is serially mediated by perceived naturalness and social judgments. We postulate that perceived naturalness alone is insufficient to create the positive effect on attitude, as supported by the findings that people quickly adapt to unnaturally sound voices (Dickerson et al., 2006). However, perceived naturalness can affect user perceptions of the agents. Perceived naturalness of VAs' voices affected perceptions and evaluation of VAs (Kühne et al., 2020), reduced perceived VA artificiality (Guha et al., 2022), and was associated with perceived conscientiousness and agreeableness of the speaker (Kühne et al., 2020). Thus, we propose that a human (vs. synthetic) voice of VAs sound more natural, which then facilitates positive social perception (i.e., competence and warmth) of VAs and consequently improves attitudes toward VA.

H3. The human voice effect will be explained by perceived naturalness of the voice, which in turn increases warmth and competence perception of the VAs; that is, perceived naturalness of the voice and warmth/competence will serially mediate the effect of voice type on attitude.

2.3. Joint effects of task type and voice

As the VAs are widely used for various activities in daily life, the nature of the tasks VAs performs can vary greatly. For example, VAs can assist consumers by quickly finding information from the internet. They can also entertain consumers by playing music or challenging them with trivia games. They can even provide companionship by engaging in conversations. Generally, the tasks consumers use the VAs can meaningfully be categorized into two types based on whether the tasks are clearly goal driven or not. Goal-oriented tasks or functional tasks serve specific purposes to accurately and efficiently solve problems at hand (e.g., finding the fastest way to the hospital, locating the nearest gas station, adding numbers). On the other hand, social tasks serve social or emotional goals and are relational in its nature (e.g., jokes, small talks) (Chattaraman et al., 2019; Whang & Im, 2021).

While the majority of consumer and marketing studies on user perception of non-human technology usage focused on the functional task context exclusively, evidence is growing in recent studies that there is a need to investigate consumer perceptions in different usage contexts because consumers respond in a context-dependent way (Dou et al., 2020; Salem et al., 2013; Schreibelmayr & Mara, 2022; Torre & White, 2021; Whang & Im, 2021). For example, Salem et al. (2013) found that people changed their perception and evaluation of robots when they interacted with the robot for different reasons (i.e., goal-oriented interactions vs. chitchats). Dou et al. (2020) similarly reported that the effects of different voices on social robot evaluations were dependent on the usage context (i.e., shopping reception, home companion,

education).

Previous findings suggest that non-human agents are better received when they perform functional tasks than social tasks. It is deemed that machines and robots can successfully assist human users in completing functional tasks which can be accomplished by locating, collecting, comparing, or analyzing facts. However, social tasks which aim to foster exchange of social information and emotions demand complex human knowledge and intuitions, and thus viewed as tasks that non-human agents do not perform well (Chattaraman et al., 2019). As reasoned in the previous section, when VAs use human voice, the voice will be perceived as more natural and thus consumers will humanize VAs. Indeed, previous research showed that people associate perceived naturalness of the voice with more humanlike abilities and competence (Torre & White, 2021). The positive effect of human (vs. synthetic) voice then will be more pronounced when the VAs are used for tasks humans can clearly perform better than machines (i.e., social tasks). Conversely, because non-human agents can perform functional tasks without difficulty, the positive effect of human voice will be diminished when consumers use VAs for functional tasks.

When evaluating the proposed psychological process and considering the two social perception mediators (i.e., warmth and competence) for different task types, it is also possible to reason that the task type moderates the human voice effect on attitude. We anticipate that perception of warmth and competence differently influence attitude toward VAs depending on the task type. Because of the different goals each task type aim to achieve, consumers are likely to rely more on competence than warmth perception for functional tasks. In contrast, warmth perception is likely to be more diagnostic than competence perception for social tasks. A recent study of service frontline robots corroborates this hypothesis. The study reported that perceived competence most strongly influenced consumers' expected utilitarian values (i.e., functional, monetary) whereas perceived warmth most strongly affected relational values (i.e., emotional) (Belanche et al., 2021). Formally, H4 is developed.

H4. The human voice effect will be diminished (amplified) when the tasks are functional (social). In other words, the task type will moderate the voice effect on attitude.

Fig. 1 visualizes the hypotheses comprehensively. Two online experiments were conducted to test the hypotheses. The next section will present the design, experiment procedures, and findings from each study.

3. Study 1

3.1. Pretests

To select functional and social tasks, a pretest was conducted prior to Study 1. 54 Amazon MTurk workers reviewed and evaluated a list of 26 common VA tasks on a 7-point scale (socially-oriented - task-oriented). The tasks were ranked based on the score from the most social (least functional) to least social (most functional) task. Five tasks that best represent each category were then selected. Using the selected 10 tasks, stimuli manipulating the voice type were developed. Two sets of audio clips were developed. To simulate the real consumer experience, we selected a widely accepted VA, Google Assistant, for the synthetic voice. To create the audio clips, a human user interacted with Google Assistant using the selected 10 tasks (e.g., "Ok, Google. What are the synonyms of diligent?"). The responses of Google Assistant were recorded. For the human voice, a female recorded the same responses. A separate pretest ($n = 61$, Australian consumers) was conducted to confirm that the human voice was perceived as more humanlike than the synthetic voice ($F = 17.743$, $p < .001$). The two voice types did not differ in terms of clarity ($p = .521$), femininity ($p = .373$), speed ($p = .206$), and fluency ($p = .078$). The task type (functional vs. social) did not cause any difference in perception of VA's responses (humanlikeness, $p = .318$; clarity, $p = .357$; femininity, $p = .101$; speed, $p = .847$; fluency, $p = .382$).

3.2. Participants

Our target sample size for Study 1 was 245 respondents. In total, 350 respondents were recruited from a professional consumer panel but 105 respondents were excluded as they: were incomplete responses ($n = 5$); failed attention checks (e.g., please select "strongly disagree" if you are paying attention - this is an attention check; $n = 9$); deemed to be straightliners ($n = 2$); and/or deemed to be bot responses by the professional consumer panel ($n = 89$ from two duplicate IPs). After excluding low-quality data, Study 1 consisted of 245 Australian adult consumers who are fluent in English and speak English daily were recruited from a professional consumer panel. 121 (49.4%) of the participants were females and 124 (50.6%) participants were males. The average age of the participants was 39 ($SD = 12$) with a range of 18–60 years. In our data analysis, we tested whether the results were affected by age, gender, and ethnicity. When entered as covariates, none of the demographic factors had a significant effect.

3.3. Design and procedure

For the main study, a 2 (between-subjects factor: human vs. synthetic voice) x 2 (within-subject factor: functional vs. social tasks) mixed design was employed using an online experiment with a self-complete online questionnaire. Participants were randomly assigned to one of the two voice type conditions and asked to listen to 10 interactions (5 functional and 5 social tasks) between a user and a VA. Because the participants were asked to imagine interacting with a VA themselves, the commands of the user were presented as texts on the screen whereas the response of the VA was presented as an audio clip. The 10 tasks were presented in a randomized order and the participants evaluated the interaction and the VA's responses after listening to the VA's response to each task.

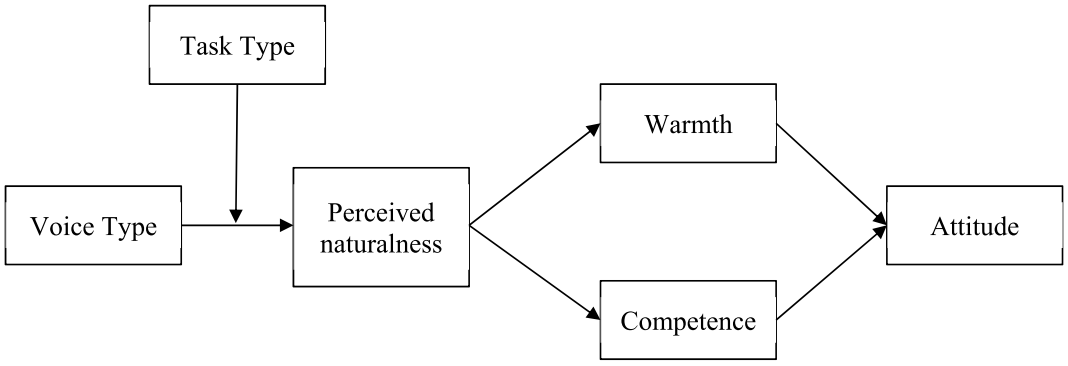

Fig. 1. Conceptual model.

For the evaluation of the interaction, measures were adopted from previous studies whenever possible. Perceived warmth (α = 0.86) and competence (α = 0.92) were measured by adopted scales from Fiske et al.'s research (2002). Attitude was measured by an item (1 = bad; 9 = good) adopted from Petty et al. (1983)'s attitude scale. Attitude was measured with a single item as it is less monotonous and time-consuming, considering that respondents evaluated 10 interactions with the VA. Furthermore, attitude has been largely considered as a unidimensional construct that can be measured with single-item measures in a valid and reliable manner (for a review, see Rossiter, 2011). This notion is also supported by recent meta-meta-analyses from Ang and Eisend (2018), which validated the use of single-item measures of attitude. We measured perceived naturalness using three 9-point bipolar items developed for this study (natural – unnatural; machine-like – human-like; very much artificial – not at all artificial (α = 0.84 in the pretest)). Lastly, a manipulation check was conducted after each interaction by asking respondents to report on a bipolar scale as to whether they perceive the interaction to be socially orientated (1) to task-orientated (9). The stimuli, survey questions, dataset, and the main output can be accessed via: https://osf.io/e5p79/?view_only=64765668820b4354922aa1d6902c1a8d.

3.4. Data analysis & results

The manipulation check ("the task was (1)socially-oriented – (9)task-oriented.") shows that respondents perceived functional tasks to be more task-orientated than social tasks, and social tasks to be more socially oriented ($M_{functional}$ = 5.61(SD = 1.61) vs. M_{social} = 3.77(SD = 1.86), t (234) = 12.33, $p < .001$). All respondents were able to correctly identify whether the voice they heard was a human or synthetic voice. Thus, the manipulations were successful. The inter-item reliability of all measures was confirmed (Cronbach's α > 0.81). The multi-item measures were then averaged for each variable.

A 2 (within-subject: social vs. functional task) x 2 (between-subjects: Human vs. synthetic voice) mixed ANOVA were conducted. Partial eta squared (partial η^2) was used as the metric of effect size, whereby a value of 0.01 denotes a small effect, 0.06 denotes a medium effect, and 0.14 indicates a large effect (Cohen, 1988). The results confirmed significant main effects of both the voice type and task type on attitude. While there was no hypothesis for the main effect of the task type, the results revealed a significantly higher positive attitude toward functional (M = 7.32, SE = 0.10) than social tasks (M = 6.07, SE = 0.13) (F (1, 243) = 4.15, p = .043, *partial* η^2 = *.02*). There was also a significant main effect of voice type as respondents showed significantly more positive attitude towards the synthetic (M = 7.04, SE = 0.15) than human voice (M = 6.34, SE = 0.15) (F (1, 243) = 3928.32, $p < .001$, *partial* η^2 = *.04*). This is opposite to our H1 that consumers would prefer human voice. Thus, H1 is rejected. Although not formally hypothesized, the voice type x task type interaction effect on attitude towards VA was also significant (F (1, 243) = 4.15, p = .017, *partial* η^2 = *.02*). The post-hoc mean difference evaluation revealed that the significant interaction effect was explained by the fact that participants rated VA with the synthetic voice (M = 7.78, SE = 0.15) more positively than human voice (M = 6.86, SE = 0.15) when engaging in functional tasks (t = 4.39, $p < .001$) but not in social tasks (t = 1.82, p = .062).

To test the mediation and moderated mediation hypotheses (H2 to H4), PROCESS procedure was used (for a review, see Hayes, 2017). In the PROCESS procedure, significant moderation and mediation is denoted by a significant 95% confidence interval (CI) bootstrapped based on 10,0000 samples (the result is significant if the range of the boostrapped CI does not include the value of zero). R^2 is used as a measure of effect size (see Table 1). The serial mediation of perceived naturalness, warmth and competence was tested. Because the task type was a within-subject factor, the moderator could not be included in the PROCESS analysis model. Therefore, two serial mediation analyses, one for functional and the other for social tasks, were conducted and compared to examine the moderating effect of task type (Model 6, bootstrap samples = 10,000).

For functional tasks, attitude towards VA was predicted by the voice type ($p < .001$) yet in the opposite direction (total effect = −0.27, 95% CI = −0.547;-0.010). This result confirms the ANOVA result, rejecting H1. H2 posited that the voice effect would be explained by perceived warmth and competence. Perceived competence was a significant mediator of the voice type effect on attitude (effect = −0.31, 95% CI = −0.530;-0.099) whereas perceived warmth was not (effect = 0.00, 95% CI = −0.037; 0.042). While competence was identified as a significant mediator, because the synthetic voice, rather than human voice, increased attitude towards VAs, this result rejects H2. Next, we hypothesized perceived naturalness as the key mediator for the process (H3). However, perceived naturalness of voice was not predicted by the voice type (p = .740) and consequently it did not explain the voice type effect on attitude. Thus, no serial mediation was significant. Thus, H3 was rejected. The direct effect of the voice type on attitude remained significant in the serial mediation model (effect = −0.65, $p < .001$, 95% CI = −0.970; −0.325), implying that the synthetic voice increased attitude towards VA partially through enhanced perceived competence.

We repeated the same analysis for social tasks. Contrasting the results from functional tasks, the voice type did not predict attitude towards VA (coefficient = −0.28, p = .114), further rejecting H1. Although all mediating variables, perceived naturalness (coefficient = 0.27, $p <$ .010), competence (coefficient = 0.59, $p < .001$), and warmth (coefficient = 0.16, $p < .050$) were positively associated with attitude, a further examination of the indirect effects revealed that no mediation of these variables was significant, rejecting H2 and H3. The direct effect of the voice type on attitude also was not significant in the serial mediation model (effect = −0.28, p = .114, 95% CI = −0.627; 0.068). Together, the results imply that the voice type does not affect consumers' attitude towards VAs for social tasks.

When comparing the results of social tasks with those of functional tasks, the task type affected perceptions of VAs and attitudes towards

Table 1
PROCESS results of Study 1 and 2.

	Study 1		Study 2	
	Functional	Social	Functional	Social
R^2	.45	.57	.85	.79
Voice → Attitude	**−.270 (-.547;-.010)**	−.204 (−.584.192)	**.511 (.055;.959)**	.149 (−.368.540)
Voice → naturalness → attitude	.015 (−.080.106)	−.029 (−.143.071)	*Not Included*	*Not Included*
Voice → warmth→ attitude	.001 (−.0366.0417)	−.007 (-.061.058)	*Not Included*	*Not Included*
Voice → competence→ attitude	**−.311 (-.530;-.099)**	−.095 (−.328.120)	−.034 (−.300.226)	−.152 (−.357.152)
Voice → naturalness → competence → attitude	.018 (−.088.123)	−.044 (−.192.103)	*Not Included*	*Not Included*
Voice → naturalness → warmth → attitude	−.003 (−.034.022)	−.095 (−.052.023)	*Not Included*	*Not Included*
Voice → fluency → attitude	*Not Included*	*Not Included*	**.094 (.010;.230)**	.063 (−.017.178)
Voice → fluency → competence → attitude	*Not Included*	*Not Included*	**.450 (.152;.732)**	.187 (−.054.420)

Note. Each cell displays the corresponding effect estimate and 95% confidence interval in the parenthesis. Statistically significant effects are denoted with bold font. The coefficient values are positive in Study 2 because the dummy coding of voice type was reversed (1 = synthetic; 0 = human).

VAs, confirming the moderating role of the task type. However, the nature of the moderation was inconsistent with our hypothesis (H4). Contrary to our prediction, the synthetic voice outperformed the human voice and this effect was limited to functional tasks. The positive effect of the synthetic voice was partially explained by enhanced perceived competence of the VA. Thus, all our hypotheses were rejected. See Table 1 for summarized hypothesis testing results.

3.5. Discussion

The results of Study 1 are unexpected and contradict the previous literature that documented user preference of human voice to synthetic voice (Baird et al., 2018; Chérif & Lemoine, 2019; Kühne et al., 2020; Schreibelmayr & Mara, 2022). While our findings seem surprising, a few studies provide insights to explain the results. It is possible that the synthetic voice used in this study is natural enough to suppress the negative impact (and maximize the positive impact) of synthetic voice. In a recent study that investigated the congruence between the realism of visual appearance and the realism of voice of virtual characters (Higgins et al., 2022), the authors did not find supporting evidence that the synthetic voice led to negative perceptions and emotional responses in their initial study in which they contrasted a natural human voice with a high-quality synthetic voice. When they introduced a lower quality synthetic voice (i.e., unnatural text-to-speech voice) in their second study, the hypothesized negative effects of synthetic voice were observed. Similarly, Craig et al. (2019) who compared the effects of old text-to-speech, modern text-to-speech, and human voice in education settings found that only the old text-to-speech was rated significantly worse than human voice, whereas modern text-to-speech and human voices were similarly effective. When looking at the synthetic voices used in previous studies, researchers who found voice naturalness as an important trait tended to manipulate the voice to clearly make it sound mechanical (e.g., McGinn & Torre, 2019). Because our study used the voice of Google Assistant which is far more realistic and natural than the traditional text-to-speech synthetic voice, it is possible that the "synthetic" voice in our study failed to produce the hypothesized effects. Given that we found no effect of voice type on perceived naturalness, the advanced text-to-speech technologies might have generated synthetic voices with sufficient naturalness perception and weakened the merits of human voices, and perceived naturalness does not explain the voice type effect.

Our results also revealed a positive effect of the synthetic voice on attitude for functional tasks. This finding is in line with a recent study which tested and confirmed consumer beliefs that computers or AI are more competent than humans when consumers have utilitarian goals (Longoni & Cian, 2022). In this study, the authors demonstrated that people believe AI outperforms humans in making recommendations for utilitarian purposes (e.g., hair products to satisfy utilitarian goals such as performance, practicality, chemical compositions) while humans are more competent than AI for hedonic purposes (e.g., hair products to satisfy hedonic goals such as scent, luxurious feeling, indulging).

As Study 1 results rejected the significance of perceived naturalness of voice, we now turn to further investigate which voice quality feature caused the users to perceive the synthetic voice as more competent for functional tasks. Previous research on social perception through vocal quality assessment revealed people use the speech rate or fluency as a sign of professional competence along with nonverbal qualities such as direct eye gaze (Leigh & Summers, 2002). In Study 2, we repeat the same experiment to confirm and validate the results of Study 1 while testing whether perceived fluency increases competence evaluation and explains the effect.

4. Study 2

4.1. Participants

Our target sample size for Study 2 was also 200 respondents. In Study 2, we ensured that participants from Study 1 and excluded samples from Study 1 were not able to participate in Study 2. In total, 210 respondents were recruited from a professional consumer panel but 17 respondents were excluded as they: were incomplete responses (n = 1); failed attention checks (n = 3); deemed to be straightliners (n = 1); and/or deemed to be bot responses by the professional consumer panel (n = 12 from 1 duplicate IP). After excluding low-quality data, there were 193 Australian adult consumers. The sample had relatively balanced gender distribution, with 93 (48.2%) respondents being female and 100 (51.8%) being males. Their average age ranged from 18 to 60 years, with a mean age of 38 years (SD = 12). All respondents must be fluent in English and speak English daily. We also considered age, gender, and ethnicity entered as covariates in our data analyses, but none of the demographic factors serve as significant covariate.

4.2. Design and procedure

The purpose of Study 2 was to test if perceived fluency of the voice is responsible for consumers' perception of enhanced competence of the VAs for functional tasks. The design and procedure of Study 2 were mostly identical to Study 1 with two exceptions. Instead of perceived naturalness, respondents were asked to evaluate the perceived fluency of the response to each task on one item: "The voice assistant's response is fluent" (1 (strongly disagree) to 7 (strongly agree) with the anchor). Because the focus of the study was on enhanced competence and the previous results suggested no voice effect on warmth, participants were asked to rate perceived competence only. This also kept the questionnaire shorter and reduced fatigue as the participants responded to the questionnaire after each of the 10 tasks. The stimuli, survey questions, dataset and the main output can be accessed via: https://osf.io/e5p79/?view_only=64765668820b4354922aa1d6902c1a8d.

4.3. Data analysis & results

The manipulation check ("the task was (1)socially-oriented – (9)task-oriented.") confirmed that respondents perceived functional tasks to be more task-orientated than social tasks, and social tasks to be more socially oriented, ($M_{func} = 6.72$ (SD = 2.02) vs. $M_{social} = 5.24$ (SD = 2.39), t (192) = 7.05, $p < .001$). All respondents were able to correctly identify whether the voice they heard was a human or synthetic voice.

A 2 (within-subjects: social vs. functional task) x 2 (between-subjects: Human vs. synthetic voice) ANOVA replicated the results of Study 1. It revealed a significant main effect of the task type, whereby respondents displayed significantly higher positive attitudes toward functional (M = 6.81, SE = 0.13) than social tasks (M = 6.61, SE = 0.12) (F (1, 191) = 7.00, p = .009, *partial* η^2 = *.04*). Consistent with Study 1 result, respondents again showed significant positive attitudes towards the synthetic (M = 6.95, SE = 0.17) than human voice (M = 6.47, SE = 0.16) (F (1, 191) = 4.38, p = .038, *partial* η^2 = *.02*). The task x voice interactive effect was also significant (F (1, 191) = 5.08, p = .025, *partial* η^2 = *.06*). Post-hoc comparison revealed that the respondents exhibited significantly higher positive attitudes towards the synthetic voice (M = 7.14, SE = 0.18) than human voice for the functional tasks (M = 6.49, SE = 0.17) (t = 2.60, p = .010), but not for the social tasks (t = 1.32, p = .182).

As in Study 1 analyses, two serial mediation analyses were conducted for each task type (PROCESS model 6, bootstrap samples = 10,000). Specifically, for the serial mediation analyses, voice type was entered as the independent variable and attitude was entered as the dependent variable. Similar to Study 1, the moderating effect of task type was examined through conducting two separate serial mediation analyses.

Perceived fluency and competence were entered as serial mediators. That is, the serial mediation analyses examine the following in a sequential or serial manner: (1) whether task type influences perceived fluency; (2) whether such differences in perceived fluency affect perceived competence; and (3) whether such differences in perceived competence lead to attitude. For functional tasks, the voice type predicted perceived fluency (coefficient = 0.51, 95% CI = 0.505; 0.959). Evaluation of mediations revealed that perceived fluency significantly mediated the direct effect of voice type on attitude (effect = 0.09, 95% CI = 0.010; 0.230). The serial mediation through fluency and competence was also significant (effect = 0.45, 95% CI = 0.1520.732). The direct effect of voice type was not significant (effect = 0.14, $t = 1.3848$, $p = .168$), suggesting that perceived fluency and competence fully mediated the voice type effect on attitude. In contrast, the voice type did not predict perceived fluency (coefficient = 0.29, $t = 1.5548$, $p = .122$) for social tasks. Furthermore, no indirect or direct effects of the voice type were significant. Therefore, no mediation was detected, replicating the results of Study 1. The analysis results are presented in Table 1.

4.4. Discussion

The results of Study 2 largely replicated those of Study 1. Consistent with Study 1, the participants rated VA with the synthetic voice more positively than the human voice only when they engage in functional tasks. Importantly, perceived fluency of the voice was confirmed as an important mediator that explains the voice type effect on attitude. The serial mediation analyses supported that perceived fluency of the voice and competence serially and fully mediated the effect of voice type on attitude when participants engage in functional tasks. This mediation effect disappeared for social tasks. Fig. 2 below visually summarizes the findings from the two studies.

5. General discussion

5.1. Summary of findings

As technological advancement in artificial intelligence enables machine agents to imitate human abilities increasingly well, VAs are getting smarter and widely available for various tasks. Today, consumers encounter and interact with VAs daily when they use their phones or smart home devices. When it comes to the voice of virtual agents and robots, the majority of literature so far advocated for the use of natural humanlike voice (Dou et al., 2020; Fan et al., 2016; Schreibelmayr & Mara, 2022). However, the exact underlying mechanism for why users prefer humanlike voice remains unclear, which limits our understanding of subtle nuances of the humanlike voice effects. To bridge this gap, this study aimed to uncover the psychological process that explains how the perception of voice influences consumers' enhanced attitude towards VAs and examine a boundary condition of the voice effect (i.e., task type). Surprisingly, results from two studies collectively rejected the widely accepted notion that human voice is more likable than synthetic voice. Instead, participants of both Study 1 and 2 exhibited more favorable attitudes towards VAs with synthetic voice than human voice, and the favorable evaluation of synthetic voice was explained by heightened perceived competence of VA when participants engaged in functional tasks. While our initial hypothesis was perceived naturalness of voice would be a key trigger for positive VA perceptions, Study 1 result rejected this hypothesis. Through Study 2, we identified perceived fluency as a mediator and demonstrated that participants perceived the VA to be more fluent when responding to commands for functional tasks, which in turn increased their judgment of VA's competence. Thus, the findings of the two studies suggest that synthetic voice enhances consumers' attitude towards VA when they engage in functional tasks because they feel that the VA responds to their command more fluently and thus believe the VA to be competent. Consumers are unaffected by the voice type when they use VAs for social tasks.

5.2. Contribution of the study

The current study is one of the first studies that investigates the effects of voice on consumer attitudes while systematically incorporating the interaction context using the social cognition perspective. Our conceptualization of the psychological process of VA-consumer interactions advances our knowledge of how the voice of VAs may influence consumers by delineating the chain of responses from the perception of voice, evaluation of VA responses, and judgment of VA. While we share some foundational assumptions with studies that tested the effects of humanlikeness of virtual agents or robots (Belanche et al., 2021; Go & Sundar, 2019; Li & Suh, 2021; Whang & Im, 2021), our study is original in that we specifically probed features of voice and how

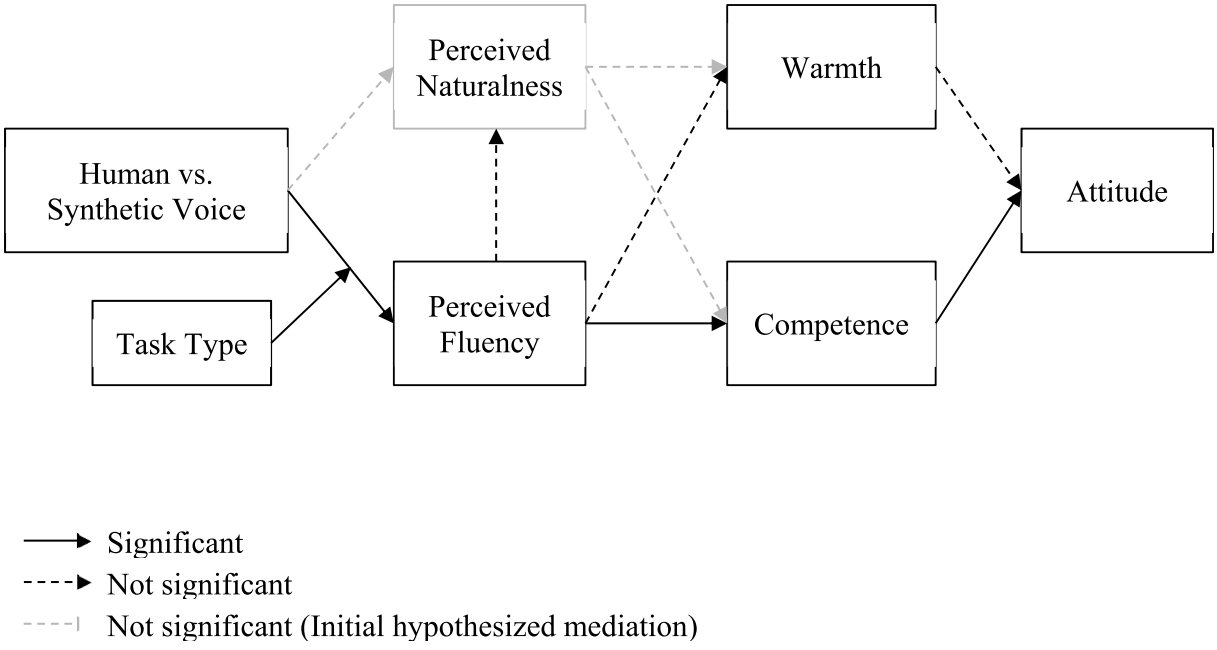

Fig. 2. Final research model integrating study 1's and 2's findings.

perceptions of vocal and speech interaction elements trigger the psychological downstream process. Although previous studies often focused on visual design features (Belanche et al., 2021; Go & Sundar, 2019), our study emphasizes the need to attend to voice design features in understanding consumers' internal responses to VAs.

Our study also adds to the growing evidence that researchers and marketers must factor in the interaction goals and contexts when designing and implementing VAs. We discovered that consumers perceive the VA's responses differently depending on the type of tasks they are engaged in. Our findings support studies that illuminated context-dependent user behaviors when interacting with virtual agents or robots (Dou et al., 2020; Higgins et al., 2022; Longoni & Cian, 2022; Schreibelmayr & Mara, 2022) and highlight the importance of investigating consumer behaviors with clear consideration of usage context.

Our study also challenges the common belief that more humanlike voice is better for VAs. The result that VA's synthetic voice positively influenced participants for functional tasks was inconsistent with previous research that documented a human voice superiority (Chérif & Lemoine, 2019; Kühne et al., 2020; Schreibelmayr & Mara, 2022). However, our finding is closely aligned with a recent study which confirmed that people exhibit different preferences between artificial intelligence (AI) and humans depending on whether their focal interest was hedonic or utilitarian (Longoni & Cian, 2022). The authors revealed that the salient importance of utilitarian (i.e., functional) attributes of a product or a service makes consumers prefer non-human AI recommenders to human recommenders because of their lay belief that AI is more competent in utilitarian tasks than humans. When applying this underlying logic to the VA interaction context, our finding directly confirms this lay belief (i.e., the significant indirect effect of synthetic voice through perceived competence). Our result also extends Longoni and Cian's study by uncovering an additional predictor, the perception of VA voice and responses that precedes competence judgment (i.e., serial mediation of perceived fluency and competence). We will discuss more about this point in the following theoretical implications section.

5.3. Theoretical implications

The current study views VAs as a humanlike entity that people treat as a social interaction partner. Drawing from the CASA framework which focuses on interactions and social perceptions, we developed the theoretical model that people use the voice of VAs to infer personality traits as they do in other human-to-human interactions. Furthermore, by incorporating the widely accepted warmth-competence evaluation of social perception, our theoretical model conceptually endorses the view that users of VAs activate certain stereotypes based on the warmth and competence judgments. While our specific hypotheses were rejected, confirmation of the serial mediation of voice quality perception – competence – attitude validates the usefulness of the CASA framework and application of social cognition to VAs. Our study provides additional support for the robust CASA literature that demonstrated the applicability of social cognition principles in machine perception (Belanche et al., 2021; Demeure et al., 2011; Gambino et al., 2020; Lee & Nass, 2010).

Our study implicitly shares the view of the stream of literature that advocates for positive effects of humanlikeness in device and robot designs. While many of these studies were focused on understanding the degree of anthropomorphism and whether hyper-realistic machines would be accepted by the users (i.e., uncanny valley hypothesis), our focal interest did not reside in the degree of humanlikeness per se. Rather, we were interested in the downstream psychological process of social perception which could be facilitated by humanlikeness of the voice. Therefore, although our hypothesis rejected the positive effects of humanlikeness (i.e., human vs. synthetic voice effect), our finding may not be simply interpreted as contradictory evidence to positive effects of humanlikeness. As previously discussed (section 3.5.), it is possible that the humanlikeness effects are not linear but logarithmic and the effects may diminish as the realism increases. Another possibility is that the humanlikeness inferred from voice alone may not produce as strong effects as visual assessment from physical appearance or combined evaluation using multiple senses (e.g., appearance, gesture, facial expression). It is also noteworthy that the interactions used in this study were very short, consisting of one-sentence responses to simple commands. Therefore, the stimulus voices might have conveyed limited social information, minimizing the potential positive effects of human voice.

In our study, perceived fluency of VA responses was identified as an important predictor. People associate verbal fluency with high intelligence and confidence (Amick et al., 2017). Therefore, it is reasonable that perceived fluency is an important underlying basis for competence judgment. It is noteworthy that pretests showed that synthetic and human voice were perceived equally fluent ($M_{synthetic} = 3.969$ vs. $M_{Human} = 3.946$, $F = 0.018$, $p = .893$). Only when the participants in the main study imagined themselves interacting with the VAs using various commands, did the different fluency perception appear between the voice type. Thus, the perception of high fluency of the responses is likely to be formulated through consumers' expectations of how VA would respond to their commands. This finding can be explained by the stereotyping- or expectancy-induced bias. Existing beliefs or stereotypes affect various aspects of cognition including information search and encoding, attention, and memory (Meppelink et al., 2019; Sherman et al., 2009). Importantly, prior beliefs even influence sensory perceptions (e.g., distortion of length perception) (Wason & Kosviner, 1966) because they extend their influence to the perception neurons (Sohn et al., 2019). Previous research (Longoni & Cian, 2022; Schreibelmayr & Mara, 2022) supported that users believe that machines and AI can be extremely competent in the functional tasks. Such a belief or a schema of computer/AI will create an expectation that VAs will perform very efficiently and respond to the users well for functional tasks, which may bias their perception of performance. When a VA responds to users with a reasonable speed and fluency in functional tasks, the users may then perceive the responses to be very fluent as the experience is consistent with their expectation. However, when a VA responds to social tasks at which people do not expect VAs to perform well, the users may perceive the responses to be not as fluent.

5.4. Practical implications

The findings of this study also provide several practical implications for VA designers and businesses. The insignificant difference in perception of VA responses, judgments of VA, and attitude towards VA for social tasks may signify that the current technology creates smooth and natural enough interactions for consumers although the interactions may not be as good as the real human. The participants in our study were able to clearly distinguish the synthetic voice from the natural human voice. Still, they liked the synthetic voice better than the human voice for functional tasks. It may be that, once VAs can respond to the users past a certain threshold of humanlikeness, humanlikeness does not contribute to enhance consumer experiences. However, the general feeling of how suitable or acceptable VAs are for a given task may inform how individuals respond and evaluate VAs. Therefore, rather than trying to simulate perfectly natural human-to-human interactions, it may be advisable to evaluate how much consumers believe VAs can perform the given tasks. Considering VA usage for business is mostly functional, our results suggest that VAs can be not only extremely effective in communicating and assisting consumers in their shopping process but also perceived as better at the tasks than humans. While we did not test the effect in different product categories, the same logic can be extended to postulate similar effects. Search products (i.e., products that can be assessed relatively easily with product specifications and facts without using them), as opposed to experience products (i.e., products that can only be properly assessed after actual usage and experience), can be something machine or AI can effectively compare and evaluate and even

outperform humans. Therefore, people may respond positively to VAs when they shop for search goods (vs. experience goods). Consistent with this logic, Xie et al. (2022) confirmed that consumers show less avoidance of machine recommendations for search (vs. experience) products.

In understanding and applying our results to voice design, it is also important to acknowledge that the physical design of the devices may play some roles in consumer perception of voice. We presented a photo of a smart speaker when participants interacted with the audio clips. The current popular devices typically take a very simple machine form (e.g., a small speaker like Google assistant or Amazon echo) and the voice is disembodied. Studies in human-robot interactions have shown that users prefer robots when humanlikeness of their physical appearance and voice is matched (McGinn & Torre, 2019). Thus, for the current design of VAs, signaling the machinelike characteristics may be beneficial because they holistically communicate a congruent level of humanlikeness. This finding may imply that VA designers should consider the device design when developing an appropriate voice for the VA.

Our study was built on the social cognition and CASA framework principles. The findings suggest that the basic premise that humans naturally treat VAs as if they are other (human) social interaction partners still holds true. The participants used the voice as the social cue to assess whether the VA is competent in performing the intended tasks just like how they would do so with other human beings. While this is beyond the scope of the current study, this implies that consumers may apply other social principles (e.g., reciprocity) when interacting with VAs. For example, people behaved politely to computers and were flattered by computer responses (Reeves & Nass, 1996). It is important that VA designers and business leaders are aware of this automatic human response to non-human machines and consider how to ethically design the VAs and interactions.

6. Limitations and future research suggestions

We note a few limitations of the current study which also suggest avenues for future research. First, we aimed to achieve realism in our study by using a currently available VA in our experiment. As a result, we only investigated the voice of an adult female to match the most widely used default voice of Google Assistant. We acknowledge that there are subtle nuances and differences in voice traits (e.g., gender, age, accent) which may affect consumer responses to the voice. Future research that expands the current study and investigates other voice features may provide additional insights. Second, our study only investigated the design of the common VAs today – the stand-alone speaker form. However, VAs are increasingly embedded in various smart devices and some VAs have displays (e.g., Amazon Echo Show). The display can be used to give humanlike physical features to the currently disembodied voice of the VAs, and future research is needed to understand how the visuals may facilitate or hinder consumer experiences. Third, when considering the direct effect of task type, the effect sizes observed in the current research are small-to-medium and this is potentially due to the nuisance of the experimental manipulation and the artificial nature of an experimental design. However, it is important to note that the serial or indirect effects of task type accounted for approximately 45%–85% of variance in user attitude toward the VA. This suggests that it is important to consider the effect of task type in the context of significant psychological mediators such as perceived fluency and competence. Nevertheless, future studies should use the current findings as a reference and examine the significant effects in a large-scale field study to ensure the effect is robust and meaningful for the design of the VAs. Fourth, our current line of research can be greatly advanced by further categorizing and differing the consumer interaction tasks. Our study made a broad distinction and contrasted two types of tasks. Although this distinction provides initial insights, this dichotomization is unsuitable to fully investigate complex and nuanced differences among various tasks. Because our finding implies the tasks consumers engage in and different levels of performance expectation form the basis for the overall attitude towards VAs, future research is necessary to identify and characterize VA tasks while considering the general beliefs about human and AI abilities. Such research will advance our understanding and allow us to produce more fine-grained practical implications for VA applications. Fourth, although we reasoned and interpreted the finding that synthetic voice enhanced perceived fluency of responses, the current study did not directly manipulate and test this underlying mechanism. Future research is necessary to validate this deduced interpretation. Lastly, it is worth pointing out that the simulated interactions with VAs in the study may have weakened the true effects of the voice. In the current study, the role of warmth perception was minimized possibly because of the nature of VA usages. However, this result may be due to the design of the study based on the short, imagined interactions. All tasks used in the current study including social tasks were very short and VAs responded to a single input with a single response. Moreover, the participants did not actually speak to the VAs but imagined their talking to the machine. In these short simulated interactions, the social and emotional goals may be harder to address than functional goals. Future studies that closely simulate the actual use of VAs and assess objective measures of user responses in addition to self-reported experience (e.g., observed behaviors, neuro-physiological responses) can further our understanding. Additionally, studies that vary the length of interactions may produce more varied degrees of user perception of warmth from the interactions with VAs.

7. Conclusion

Consumers use VAs daily for various tasks. The current study investigated consumer perceptions of VAs based on their typical interactions with VAs. Our findings clearly showed that consumers neither anticipate nor prefer extremely humanlike voice from VAs. Instead, consumers value VAs for their ability to perform functional tasks effectively and form positive attitudes. Efforts to humanize the VAs could weaken this effect. As we begin to understand how average consumers interact with VAs, the findings of this study provide a valuable insight. While designers and developers strive to create a machine that can speak just like humans and make innovative products, the users for the VA applications prefer the VAs with reasonably machinelike voice, and think VAs are highly competent for the tasks they believe VAs would do well. Understanding and managing consumer expectations and using the knowledge to create VA designs that match consumer expectations may yield far more successful results.

Credit author statement

Hyunjoo Im: Conceptualization, Methodology, Writing – original draft preparation, Writing – review & editing, **Billy Sung**: Conceptualization, Methodology, Formal analysis, Investigation, Supervision, Writing – original draft preparation, Writing – review & editing, **Garim Lee**: Writing – original draft preparation, Writing – review & editing, **Keegan Qi Xian Kok**: Formal analysis, Investigation, Writing – original draft preparation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The stimuli, survey questions, dataset and the main output can be accessed via: https://osf.io/e5p79/?view_only=64765668820b4354922aa1d6902c1a8d.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2023.107791.

References

Amick, L. J., Chang, S.-E., Wade, J., & McAuley, J. D. (2017). Social and cognitive impressions of adults who do and do not stutter based on listeners' perceptions of read-speech samples. *Frontiers in Psychology, 8*, 1148. https://doi.org/10.3389/fpsyg.2017.01148

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

Belanche, D., Casaló, L. V., Schepers, J., & Flavián, C. (2021). Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model. *Psychology and Marketing, 38*(12), 2357–2376. https://doi.org/10.1002/mar.21532

Cambre, J., & Kulkarni, C. (2019). One voice fits all?: Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–19. https://doi.org/10.1145/3359325. CSCW).

Carolus, A., Wienrich, C., Törke, A., Friedel, T., Schwietering, C., & Sperzel, M. (2021). 'Alexa, I feel for you!' Observers' empathetic reactions towards a conversational agent. *Frontiers of Computer Science, 3*, 46. https://doi.org/10.3389/fcomp.2021.682982

Chattaraman, V., Kwon, W.-S., Gilbert, J. E., & Ross, K. (2019). Should AI-based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior, 90*, 315–330. https://doi.org/10.1016/j.chb.2018.08.048

Chérif, E., & Lemoine, J.-F. (2019). Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing, 34*(1), 28–47. https://doi.org/10.1177/2051570719829432

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Core dna. (2021). *#ChatbotFail: 4 chatbot customer experience fails uncovered*. Core DNA. May 2 https://www.coredna.com/blogs/chatbot-fail.

Craig, S. D., Chiou, E. K., & Schroeder, N. L. (2019). The impact of virtual human voice on learner trust. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 63*(1), 2272–2276. https://doi.org/10.1177/1071181319631517

Demeure, V., Niewiadomski, R., & Pelachaud, C. (2011). How is believability of a virtual agent related to warmth, competence, personification, and embodiment? *Presence: Teleoperators and Virtual Environments, 20*(5), 431–448. https://doi.org/10.1162/PRES_a_00065

Dickerson, R., Johnsen, K., Raij, A., Lok, B., Stevens, A., Bernard, T., & Lind, D. S. (2006). Virtual patients: Assessment of synthesized versus recorded speech. *Studies in Health Technology and Informatics, 119*, 114–119.

Dou, X., Wu, C.-F., Lin, K.-C., Gan, S., & Tseng, T.-M. (2020). Effects of different types of social robot voices on affective evaluations in different application fields. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00654-9

Fan, A., Wu, L., Laurie), & Mattila, A. S. (2016). Does anthropomorphism influence customers' switching intentions in the self-service technology failure context? *Journal of Services Marketing, 30*(7), 713–723. https://doi.org/10.1108/JSM-07-2015-0225

Fernandes, T., & Oliveira, E. (2021). Understanding consumers' acceptance of automated technologies in service encounters: Drivers of digital voice assistants adoption. *Journal of Business Research, 122*, 180–191. https://doi.org/10.1016/j.jbusres.2020.08.058

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878. https://doi.org/10.1037/0022-3514.82.6.878

Fraj, S., Grenez, F., & Schoentgen, J. (2009). Perceived naturalness of a synthesizer of disordered voices, 2009 *Interspeech*, 2907–2910. https://doi.org/10.21437/Interspeech.2009-736.

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–86. https://doi.org/10.30658/hmc.1.5

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Guha, A., Bressgott, T., Grewal, D., Mahr, D., Wetzels, M., & Schweiger, E. (2022). How artificiality and intelligence affect voice assistant evaluations. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-022-00874-7

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Higgins, D., Zibrek, K., Cabral, J., Egan, D., & McDonnell, R. (2022). Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics, 104*, 116–128. https://doi.org/10.1016/j.cag.2022.03.009

Imhof, M. (2010). Listening to voices and judging people. *International Journal of Listening, 24*(1), 19–33. https://doi.org/10.1080/10904010903466295

Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. *Marketing Letters, 30*(1), 1–12. https://doi.org/10.1007/s11002-019-09485-9

Krenn, B., Schreitter, S., & Neubarth, F. (2017). Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI & Society, 32*(1), 65–77. https://doi.org/10.1007/s00146-014-0569-0

Kühne, K., Fischer, M. H., & 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. *Frontiers in Neurorobotics, 14*, 105. https://doi.org/10.3389/fnbot.2020.593732

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Lee, J.-E. R., & Nass, C. (2010). Trust in computers: The Computers-Are-Social-Actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives* (pp. 1–15).

Leigh, T. W., & Summers, J. O. (2002). An initial evaluation of industrial buyers' impressions of salespersons' nonverbal cues. *Journal of Personal Selling and Sales Management, 22*(1), 41–53.

Li, M., & Suh, A. (2021). *Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.493

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. Hedonic contexts: The "word-of-machine" effect. *Journal of Marketing, 86*(1), 91–108. https://doi.org/10.1177/0022242920957347

McDonough, M. (2020). *Artificial intelligence is now shockingly good at sounding human*. December 9). Scientific American https://www.scientificamerican.com/video/artificial-intelligence-is-now-shockingly-good-at-sounding-human/.

McGinn, C., & Torre, I. (2019). *Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. 2019 14th ACM*. IEEE International Conference on Human-Robot Interaction (HRI). https://doi.org/10.1109/HRI.2019.8673305, 211–221.

Meppelink, C. S., Smit, E. G., Fransen, M. L., & Diviani, N. (2019). I was right about vaccination": Confirmation bias and health literacy in online health information seeking. *Journal of Health Communication, 24*(2), 129–140. https://doi.org/10.1080/10810730.2019.1583701

Mourey, J. A., Olson, J. G., & Yoon, C. (2017). Products as pals: Engaging with anthropomorphic products mitigates the effects of social exclusion. *Journal of Consumer Research, 44*(2), 414–431. https://doi.org/10.1093/jcr/ucx038

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181. https://doi.org/10.1037/1076–898X.7.3.171

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics, 5*(2), 171–191. https://doi.org/10.1007/s12369-012-0171-x

Nijholt, A. (2003). Disappearing computers, social actors and embodied agents. In *Proceedings. 2003 International Conference on Cyberworlds* (pp. 128–134). https://doi.org/10.1109/CYBER.2003.1253445

Number of voice assistants in use worldwide 2019-2024. (n.d.). Statista. Retrieved September 14, 2022, from https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 1*, 7–19.

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research, 10*(2), 135–146.

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology and Marketing, 38*(4), 626–642. https://doi.org/10.1002/mar.21457

PricewaterhouseCoopers. (2018). *Consumer intelligence series: Prepare for the voice revolution*. PwC. https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language and Information; Cambridge University Press.

Rella, S. (2021). *Anatomy of an AI-powered voice assistant. Speech Technology Magazine*. October 26 https://www.speechtechmag.com/Articles/ReadArticle.aspx?ArticleID=149741.

Salem, M., Ziadee, M., & Sakr, M. (2013). Effects of politeness and interaction context on perception and experience of HRI. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social robotics* (pp. 531–541). Springer International Publishing. https://doi.org/10.1007/978-3-319-02675-6_53.

Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology, 13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.787499.

Seymour, W., & Van Kleek, M. (2021). Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–16. https://doi.org/10.1145/3479515

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology, 96*(2), 305. https://doi.org/10.1037/a0013778

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron, 103*(5), 934–947.e5. https://doi.org/10.1016/j.neuron.2019.06.012

Stern, S. E., Mullennix, J. W., & Yaroslavsky, I. (2006). Persuasion and social perception of human vs. Synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies, 64*(1), 43–52. https://doi.org/10.1016/j.ijhcs.2005.07.002

Stroessner, S. J., & Benitez, J. (2019). The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics, 11*(2), 305–315. https://doi.org/10.1007/s12369-018-0502-7

Torre, I., Latupeirissa, A. B., & McGinn, C. (2020). *How context shapes the appropriateness of a robot's voice. 2020 29th IEEE International Conference on Robot and Human Interactive Communication*. RO-MAN). https://doi.org/10.1109/RO-MAN47096.2020.9223449, 215–222.

Torre, I., & White, L. (2021). Trust in vocal human–robot interaction: Implications for robot voice design. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 305–322). Springer. https://doi.org/10.1007/978-981-15-6627-1_16.

Urakami, J., Sutthithatip, S., & Moore, B. A. (2020). The effect of naturalness of voice and empathic responses on enjoyment, attitudes and motivation for interacting with a voice user interface. In M. Kurosu (Ed.), *Human-computer interaction. Multimodal and natural interaction* (pp. 244–259). Springer International Publishing. https://doi.org/10.1007/978-3-030-49062-1_17.

Vernuccio, M., Patrizi, M., & Pastore, A. (2022). Delving into brand anthropomorphisation strategies in the experiential context of name-brand voice assistants. *Journal of Consumer Behaviour*, 1–10. https://doi.org/10.1002/cb.1984

Voorveld, H. A. M., & Araujo, T. (2020). How social cues in virtual assistants influence concerns and persuasion: The role of voice and a human name. *Cyberpsychology, Behavior, and Social Networking, 23*(10), 689–696. https://doi.org/10.1089/cyber.2019.0205

Wang, W. (2017). Smartphones as social actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior, 68*, 334–344. https://doi.org/10.1016/j.chb.2016.11.022

Wason, P. C., & Kosviner, A. (1966). Perceptual distortion induced by reasoning. *British Journal of Psychology, 57*(3–4), 413–418. https://doi.org/10.1111/j.2044-8295.1966.tb01044.x

Whang, C., & Im, H. (2021). I like Your Suggestion!" the role of humanlikeness and parasocial relationship on the website versus voice shopper's perception of recommendations. *Psychology and Marketing, 38*(4), 581–595. https://doi.org/10.1002/mar.21437

Williams, R. (2019, September 19). *Study: 70% of people will swap store visits for voice assistants by 2022. Marketing Dive*. https://www.marketingdive.com/news/study-70-of-people-will-swap-store-visits-for-voice-assistants-by-2022/563238/.

Xie, Z., Yu, Y., Zhang, J., & Chen, M. (2022). The searching artificial intelligence: Consumers show less aversion to algorithm-recommended search product. *Psychology and Marketing, 39*(10), 1902–1919. https://doi.org/10.1002/mar.21706