The Effect of Naturalness of Voice and Empathic Responses on Enjoyment, Attitudes and Motivation for Interacting with a Voice User Interface

Jacqueline Urakami(✉), Sujitra Sutthithatip, and Billie Akwa Moore

Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-Ku, Tokyo 152-8552, Japan
urakami.j.aa@m.titech.ac.jp

Abstract. In human-computer interaction much attention is given to the development of natural and intuitive Voice User Interfaces (VUI). However, previous research has shown that humanlike systems will not necessarily be perceived positive by users. The study reported here examined the effect of human likeness on users' rating of enjoyment, attitudes and motivation to use VUI in a Wizard-of-Oz experiment. Two attributes of human likeness, voice of the system (humanlike vs. machinelike) and social behavior of the system (expressing empathy vs. neutral) were manipulated. Regression analyses confirmed that perceived empathy of the VUI improved interaction enjoyment, attitude towards the system, and intrinsic motivation but no effect of voice was found. Session order also affected participants' evaluation. In the second session, participants rated the VUI as more negative than in the first session. The results indicate that a VUI that expresses social behavior (e.g. showing empathy) is perceived as more favorable by the user. Furthermore, changing user expectations pose a challenge for the design of the VUI. The dynamics of user interactions must be taken into account when designing the VUI.

Keywords: Voice user interface · Empathy · Human likeness

1 Introduction

Voice User Interfaces (VUI) are widely used and are incorporated in many commercial applications such as home appliances like washing machines and dish washers, automobiles or computer operating systems. Artificial Intelligence (AI) and Deep learning has opened the door to more advanced applications of VUI such as intelligent agents that process customer requests and automated service attendants that take orders, make reservations or give directions. It is in the interest of the service provider that users can communicate with VUI in a positive way and are motivated to use the system again.

The central goal in designing VUI that act as conversational agents has been to create a natural and intuitive interaction between the user and the technical system. It is still disputed in the literature whether humans treat such systems like other social partners,

M. Kurosu (Ed.): HCII 2020, LNCS 12182, pp. 244–259, 2020.
https://doi.org/10.1007/978-3-030-49062-1_17

applying social rules and expectations to them as described in Reeves and Nass [1] media-equation hypothesis. According to this view human-computer interaction improves when the system provides many social cues and shows humanlike attributes in its behavior and appearance. In contrast the theory of human-automation assumes that humans respond to machines in a unique and specific way proposing that humans judge and evaluate machines fundamentally different [2, 3].

Goal of the study presented here is to test how users evaluate the human likeness of a VUI in regard to interaction enjoyment, attitude and intrinsic motivation. For VUI used in the service setting it is imperative that users have a positive impression of the system and are willing to engage with the system again in the future. The VUI is also seen as a representative of the company and evaluation of such a system likely will affect users' perception of the company employing it [4].

Previous studies have especially focused on human likeness in appearance of computer agents. The unique aspect of the study presented here is that the focus is not only on appearance (voice) but also considers the social behavior of the system. In contrast to other studies, particular consideration is given to how system behavior is perceived by the user and how this perception affects the evaluation of the system. This study examined two attributes of human likeness, voice of the system (humanlike vs. machinelike) and social behavior of the system (expressing empathy vs. neutral).

The results of this study highlight the importance of VUI behaving in a social way, e.g. by expressing empathy in human-computer interaction, but also suggest that individual differences and user preferences must be taken into account when designing dialogues. Furthermore, the role of changing users' expectations towards VUI is discussed and consequences for the design of VUI are contemplated.

1.1 Media-Equation vs. Human Automation Theory

The media-equation hypothesis [5, 6] implies that humans treat computers as social entities. This paradigm has been successfully studied with other technologies such as computer agents, twitter bots or robots [4, 7, 8]. The theory assumes that people react automatically and unconsciously to computers in a social and natural way, even if they believe that it is not reasonable to do so [1]. Human likeness in appearance and behavior triggers anthropomorphism, the tendency of people to attribute human-like characteristics to a computer agent. It is proposed that peoples' social reactions increase when the computer agent shows human-like behavior and provides more social cues [6, 9].

In contrast the theory of human automation proposes that human respond to machines in a unique and specific way. Research about trust showed that humans perceive the advice given by a computer agent as more objective and rational than that of a human [10]. Computer agents evoke higher initial levels of trust compared to humans, but also people lose trust in computers more dramatically [11]. Furthermore, interactions with agents that display humanlike behavior make it more difficult for people to distinguish if they are engaging with a real human or with an artificial intelligence. As research about the uncanny valley (an eerie feeling about computer agents that seem to be like humans but somewhat different) has shown, anthropomorphism can have a negative impact on peoples' evaluation of agents. This eerie feeling is especially evoked when users perceive a mismatch between a systems humanlike appearance and behavior [12].

In sum, even though human likeness triggers anthropomorphism and might increase users' social reactions toward a computer agent, it remains unclear if this is an advantage or disadvantage for human-computer interaction. Previous research has produced mixed results (e.g. [4, 13, 14]) and the question whether an automated system should pretend to be a human or act like a human is still unanswered. The goal of this study is to examine if human likeness of a VUI's voice and social behavior expressing empathy increases peoples' likeness of such systems or not.

The following two sections give a short literature overview about the role of human-like voice and empathic expressions on users' perception of automated systems.

1.2 Human-Likeness in Human-Computer Interaction

The voice is a powerful tool to create a certain impression in the listener. The voice communicates content not only through lexical expressions but also through emotional cues, and paralinguistic. In our study we focus on the human likeness of the voice compared to a machinelike voice.

Google duplex (https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html) has impressively demonstrated that it is possible to create a human-like voice with AI that is hardly distinguishable from a real human voice. When Google introduced the system, it created a controversy if it is ethical acceptable to give people the impression that they talk with a real person while they were communicating with a computer agent. This discussion makes it evident that some people feel uncomfortable interacting with computer agents that are very similar to a real person.

On the other hand, people usually rate computer agents with a natural voice more positive compared to agents with machinelike voices. Ilves [15] reported that participants rated the humanlike voice as being more pleasant and clear, and as being more approachable compared to a machinelike voice. Similarly, Louwerse [7] reported that participants preferred agents with natural voices suggesting that voice alone can trigger social behaviors towards computer agents. However Ilves [15] also noted that emotional expressions by the machinelike voice were rated by their emotional valence suggesting that machinelike voices can communicate affective messages as well.

An important factor affecting the perception of anthropomorphism seems to be how well human likeness in appearance and behavior match. Mitchell [12] paired in an experiment a robot and human face with either a machinelike or humanlike voice. Incongruence between visual appearance and voice (e.g. robot face with humanlike voice) created a feeling of eeriness. In another study Muresan [16] observed peoples interaction with a chatbot. When the chatbot violated social norms, people disapproved of the behavior and decreased the interaction, noting that the chatboot felt fake. Ruijten [8] looked at the dynamics of perceived human likeness when interacting with robots. Depending on the behavior of the robot, perceived human likeness changed over the course of the interaction. Ruijten [8] suggested that although appearance might initially be an important factor in perceiving human likeness in an agent, ultimately appropriate agent behavior, as would be expected in a particular social context, most strongly influenced the overall assessment of the agent.

Human likeness can be expressed in many ways, not only in appearance like voice but also in the form of social behavior. The next section reflects on the role of empathic responses as a form of human likeness of computer agents.

1.3 Empathy

Automated systems need to engage in a meaningful social way with people [17]. This is particularly relevant for a service provider, who needs to communicate effectively, listen actively to the problems of their customers, and provide help.

Previous research has experimented with integrating affective expressions and related responses into human-computer interactions to create a more natural humanlike experience [18, 19]. In this regard the concept of empathy has received much attention. How to define the term empathy is still an ongoing discussion in the literature. We believe that for designing interactions between humans and VUI an interpersonal approach to empathy is imperative. Therefore, empathy is defined in this paper as understanding and responding appropriately to the affective state and situation of the communication partner. Empathy is central to interpersonal functioning [20], and this paper assumes that empathic responses by a VUI can engage users in meaningful social interactions.

Empathy serves social functions when two people interact and undergoes dynamic changes in the course of this interaction, a view shared by Main [21] and Butler [20, 22]. Empathy is nested in a specific context. It is not the behavior itself that is empathic but how the behavior is suited in the context of the interaction [23]. This empathy approach emphasizes that a certain behavior must also be judged as empathic by the communication partner. It is decisive whether the user perceives and interprets the behavior of an agent as empathic.

Previous studies about the effect of empathy in human-computer interaction have reported mixed results. Złotowski [13] compared a humanlike robot with a machine-like robot and found that the humanlike robot was perceived as being less trustworthy and empathic than the machinelike robot. Niculescu [19] examined the effect of empathic language cues on users' evaluation of human-robot interaction. Even though users indicated that they perceived the robot using empathic language cues as more receptive and felt more ease in the interaction, no effect on likeability, trustworthiness, overall enjoyment and interaction quality were shown. In a study by Leite [18] an empathic model was applied to a social robot to study long-term interaction between children and robots. Leite [18] reported that the children preferred the emotion-oriented robot over the task-oriented robot. Brave and Nass [24] compared the effect of expressing self-oriented emotions and empathic-emotions with an computer agent and found that empathic-emotions resulted in higher ratings for likeability, trustworthiness, and greater perceived care and support. Araujo [4] manipulated humanlike cues in language style, name and framing in an experiment with a chatbot and reported that humanlike cues influenced the emotional connection a human felt to the agent. Urakami [25] conducted a survey study about the perception of empathic expressions by a VUI. While most participants in the study accepted empathic expressions, 1/3 of the participants disagreed with this assessment. It was concluded in the study that not all users feel comfortable with a machine acting or "pretending" to act like a human being.

Whether empathy improves human-computer interaction is still unclear. Personal preferences as well as situation adequateness of empathic expressions might affect how people react to an empathic computer agent. Empathic reactions of a computer agent must be adapted to the situational context, the role of the interaction partner and their relationship. Furthermore, it is not necessarily the empathic response itself that is decisive, but the interpretation of a certain behavior as empathic.

1.4 Hypotheses

Overall, previous research suggests that people rate a system using a humanlike voice more positive. We propose in hypothesis 1:

H1: Interaction enjoyment, attitude, and intrinsic motivation are rated higher for VUI with humanlike voice compared to machinelike voice.

Furthermore, previous studies suggest [15, 19] that emotional cues by a system will be recognized as such by users despite the naturalness or human likeness of the system.

H2: Empathic language cues by a VUI can be identified by users as an expression of empathy independently of being expressed by a humanlike or machinelike voice.

Empathy is here seen as an expression of humanlike behavior relevant in a situation where a customer approaches an agent to receive help. We propose that people would expect empathic reactions in such situations and therefore evaluate them positive for the interaction.

H3: Perceived empathy has a positive effect on users' evaluation of interaction enjoyment, attitude towards the system, and intrinsic motivation to use the system.

However, a mismatch between humanlike appearance (humanlike voice) and behavior (expressing empathy or not) will result in a negative evaluation of the agent.

H4: A VUI using a humanlike voice and empathic language cues is evaluated higher than a VUI using a humanlike voice and neutral language.

H5: A VUI using a machinelike voice and empathic language cues receives a lower evaluation than a VUI using a machinelike voice and neutral language.

These hypotheses were tested in an experimental Wizard-of-Oz study in a service setting as background scenario.

2 Method

2.1 Participants

A total of 60 international students from Tokyo Institute of Technology participated in the study, between the ages of 19 and 36 years (M = 24.95; SD = 3.72). There were 29 females and 31 males in the sample.

2.2 Experimental Design

A 2×2 mixed factorial design with within-factor empathy (expressing empathy vs. neutral) and the between factor system voice (humanlike vs. machinelike) was conducted. Two scenarios in which participants had to imagine having lost either a wallet or backpack were created. The within-factor empathy and the scenarios were counterbalanced.

2.3 Material

Scenarios. Participants had to image to have lost a wallet (scenario 1) or a backpack (scenario 2) at the airport. Participants received the following instruction: "You are at Narita airport. Your plane will depart in 1 h and you just noticed that you have lost your wallet. You used your wallet just 10 min ago to pay for extra luggage at Counter A. You need to get your wallet back as soon as possible and you ask an automated airport assistant for help. Please speak to the airport assistant: Start the conversation by saying "Hello". The backpack scenario was very similar, only that participants had to imaging they just arrived at the airport and had to catch a bus to the hotel. Both scenarios contained images of the situation and of the lost wallet and backpack.

Dialogs. The dialogs were pre-recorded using Googles text-to-speech tool (https://cloud.google.com/text-to-speech/) with the en-US-Wavenet-F voice (female voice) for the humanlike voice. The speech synthesizer Audacity (https://www.audacityteam.org/) was used to transform the voice file into a machinelike voice. By creating a specific setting in the instruction (having lost wallet/backpack at the airport) and giving the participants a specific task (try to get your lost items back) we were able to guide the participants through the dialog with the pre-recorded sequences without having the participants notice that the experiment was a wizard-of-oz study.

For each scenario (wallet/backpack) we created two dialogs with one containing empathic language and the other being neutral. Empathic phrases were chosen based on a survey conducted in a previous study by Urakami et al. [25]. The study revealed that expressions showing interest and understanding of the situation of the person, as well as expressions of offering help are being perceived as empathic by people. Table 1 displays the empathic and neutral dialog for the wallet scenario. The backpack scenario followed a similar pattern.

Negative Attitude Towards Robots Scale (NARS). The negative attitude towards robots scaly by Syrdal et al. [26] contains 14 items measuring peoples' attitudes towards situations and interactions with robots, social influence of robots, and attitude towards emotions in the interaction with robots. Items were assessed using Likert scales ranging from 1 (I do not feel anxiety at all) till 6 (I feel anxiety very strongly). The items of the questionnaire were adjusted to the purpose of the study by replacing the term robot with automated system.

CARE Measure. The CARE measure [27] is originally an instrument to measure patients' perception of empathy after the consultation with a doctor. The CARE measure consists of ten 5-point Likert scale items ranging from 1 (poor) till 5(excellent). For the purpose of our study, the term doctor was replaced by the term system. Participants evaluated if the system made them feel at ease, really listened, understood their concerns, etc.

Interaction Enjoyment. This questionnaire was adapted from Sacco and Ismail [28]. The questionnaire measures how much somebody enjoyed the interaction answering five items on 7-point Likert scales ranging from 1 (not at all) till 7 (extremely) regarding enjoying the interaction, feeling nervous, or anxious during the interaction (for Items see Table 2).

Table 1. Dialog spoken by the VUI. Sentences in *italic* are empathic expressions.

Empathic	Neutral
Hello. *Is there anything I can do for you?*	Hello. Is there a problem?
Yes, *I'd like to help you.* *Can you give me a view more details?* When did you last use it?	Yes, that is part of the provided service. When did you last use it?
Can you please tell me a bit more about it? What does your backpack look like?	Please provide more information. What does your backpack look like?
Can you please give me a few more details about it?	Please provide a few more details about it.
I am going to take care of this for you. Please wait a few seconds. I will check the airport's video footage.	OK. Please wait a few seconds. I will check the airport's video footage.
I am going to take care of this for you. Please wait a few seconds. I will check the airport's video footage.	OK. Please wait a few seconds. I will check the airport's video footage.
Thank you for waiting. *The video footage shows that someone must have stolen you backpack, but I think we can work this out.* *I can help you making a call to the police, if you would like me to do.*	The video footage shows that someone must have stolen you backpack. Should the police be called?
I'd be happy to do this for you. Thank you for using my service.	The police has been informed. Thank you for using this service.
Thank you for waiting. *The video footage shows that someone must have stolen you backpack, but I think we can work this out.* *I can help you making a call to the police, if you would like me to do.*	The video footage shows that someone must have stolen you backpack. Should the police be called?

Attitude Towards Using the System. Participant's attitude towards the system was measured using Davis et al. [29] questionnaire. Items were assessed by 7-point Likert scales ranging from 1-strongly disagree till 7-strongly agree (for Items see Table 2).

Intrinsic Motivation. This scale measured the intention of users to perform an activity with the system motivated by positive feelings while using it. The scale used Davis et al. [30] questionnaire that contained 3 items and participants indicated on 7-point Likert scales ranging from 1-strongly disagree till 7-strongly agree how much they agreed with each statement (for Items see Table 2).

Table 2. Questionnaires used in the study

Interaction enjoyment	
	I was nervous during the interaction with the system
	Interacting with the system made me anxious
	I enjoyed interacting with the system
	The interaction I had with the system was interesting
	I would enjoy interacting with the system again in the future
Attitude towards using the system	
	Using the system is a good idea
	Using the system is a foolish idea
	I like the idea of using the system
	Using the system is unpleasant
Intrinsic motivation	
	I find using the system enjoyable
	The actual process of using the system is pleasant
	I have fun using the system

2.4 Procedure

The experiment was carried out in single sessions. Participants had to attend two sessions that were scheduled at least one week apart. Participants were greeted and briefed about the purpose of the study. After signing the consent form, participants filled in the NARS. All questionnaires were presented on the computer using Google Forms. A voice sample was played to the participants to ensure that they could clearly understand the voice. If necessary, the volume of the voice was adjusted. The participants received a written instruction and when they were ready to start the experiment, they initiated the dialogue by saying "Hello" to the VUI. The participants were sitting alone on a table with the VUI in front of them. The experimenter was in the same room but hidden behind a screen. The experimenter played the voice samples according to the scripted dialog (see Table 1). Participants were given the impression that the VUI was an autonomous system that can process and react to spoken language. After the dialog was finished, participants filled in the CARE measure, interaction enjoyment questionnaire, attitude towards using the system questionnaire, and the intrinsic motivation questionnaire. After finishing the second session participants received 1000 Yen compensation.

2.5 Equipment

The humanlike voice sample used in the experiment were generated with Google cloud text-to-speech (https://cloud.google.com/text-to-speech/). The speed was set at 0.80, pitch 0.00 and the en-US-Wavenet-F voice (female voice) was used. The machinelike voice was generated using the software Audacity (https://www.audacityteam.org/). A typical characteristic of machinelike voice is being monotone and having a smaller frequency range compared to a natural human voice. Therefore, the humanlike voice samples generated with Google cloud were altered with Audacity software using first an equalizer to gradually cutting off frequencies below 200 Hz and above 800 Hz and the tempo of speech was reduced for 15%.

The VUI was a small JBL Bluetooth speaker (71.2 mm × 86 mm × 31.6 mm). The speaker was connected to the notebook of the experimenter via Bluetooth.

3 Results

3.1 Negative Attitude Towards Robots Scale NARS

Participants reported a slightly negative attitude towards automated systems ($M = 3.64$; $SD = 0.92$). There were no differences in attitude towards automated systems across the voice conditions (human: $M = 3.59$; $SD = 0.68$; machine: $M = 3.69$; $SD = 1.12$; t (58) $= -.44, p = .66$, and no gender differences, (female: $M = 3.65$; $SD = 0.948$; male: $M = 3.63$; $SD = 0.90$; t (58) $= 0.93, p = .92$).

3.2 Voice

There was no difference in being able to clearly understand the voice across the experimental conditions, t (58) $= .701, p = .486$. Also, participants could distinguish between a humanlike and a machinelike voice, t (58) $= -2.70, p = .008$.

3.3 Perception of Empathy

The CARE measure was used to control if the empathic language cues were perceived as expressions of empathy. A 2 × 2 mixed factorial ANOVA revealed a marginal significant effect of empathy (F (1, 58) $= 3.34$, $p = .07$) and no effect of voice (F (1, 58) < 1). There were no significant interactions (F (1, 58) < 1. Means and standard deviations are displayed in Table 3.

Even though the results of the CARE measure indicate that participants perceived the empathic condition indeed as empathic, the differences were rather small. As has been pointed out in the literature, empathy evolves during an interaction between two communication partners and can change during an interaction [22, 31]. It is imperative that a certain behavior really is perceived as being empathic.

To test our hypothesis about the effect of empathy on participants' evaluation of the system, we conducted first a 2×2 repeated measures ANOVA with the within-factor empathy and the between-factor voice. In addition, a regression analysis was conducted with the CARE measure as perceived empathy as a predictor.

3.4 Effect of Empathy and Voice on Interaction Enjoyment, Attitude Towards the VUI and Intrinsic Motivation

A 2×2 mixed factorial ANOVA revealed no significant effects of empathy and voice on interaction enjoyment, empathy: $F\,(1, 58) < 1$; voice $F\,(1, 58) = 1.86, p = .178$), attitude towards the VUI, empathy F (1, 58) < 1, voice F (1, 58) < 1) or intrinsic motivation, empathy $F\,(1, 58) < 1$, voice $F\,(1, 58) < 1$. Also, no significant interactions were found. See means and standard deviations in Table 3.

It was somewhat surprising that the voice had no impact on participants' evaluation, especially since the voice was clearly perceived differently in the humanlike condition compared to the machinelike condition. It can be assumed that in a direct comparison the participants might have preferred the humanlike voice. Since each participant was assigned to just one of the voice conditions, the voice itself does not seem to be a factor affecting participants overall evaluation of the system.

Table 3. Means and Standard Deviations in parenthesis for CARE measure, Interaction enjoyment, Attitude, and Intrinsic motivation

Empathy	Empathic		Neutral	
Voice	Human like	Machine like	Human like	Machine like
CARE measure	3.71 (.65)	3.74 (0.80)	3.63 (0.65)	3.53 (0.79)
Interaction enjoyment	5.46 (.88)	5.09 (0.99)	5.36 (0.82)	5.14 (0.91)
Attitude scale	5.52 (.91)	5.65 (1.12)	5.50 (1.04)	5.60 (0.89)
Intrinsic motivation	4.97 (.95)	5.14 (0 .96)	5.01 (0.99)	5.02 (0.97)

3.5 Effect of Empathy Perception on Interaction Enjoyment, Attitude, and Intrinsic Motivation

A simple linear regression analysis was used to test if the perception of empathy significantly predicted participants' ratings of interaction enjoyment, attitude, and intrinsic motivation. The CARE measurement was used as a predictor for perceived empathy. The results of the regression indicated that perception of empathy explained 34% of the variance for interaction enjoyment ($R2 = .34$, F (1, 119) $= 61.08, p < .01$). It was found that perceived empathy predicted interaction enjoyment ($\beta = .72, p < .001$).

Furthermore perceived empathy explained 36% variation for attitude ($R2 = .36$, F (1, 119) $= 67.51, p < .001, \beta = .80, p < .001$), and 33% of the variation for intrinsic motivation ($R2 = .33$, F (1, 119) $= 57.28, p < .001$, $\beta = .75, p < .001$). Figures 1, 2 and 3 display the scatter plots with regression lines for interaction enjoyment, attitude and intrinsic motivation with the CARE measure as predictor.

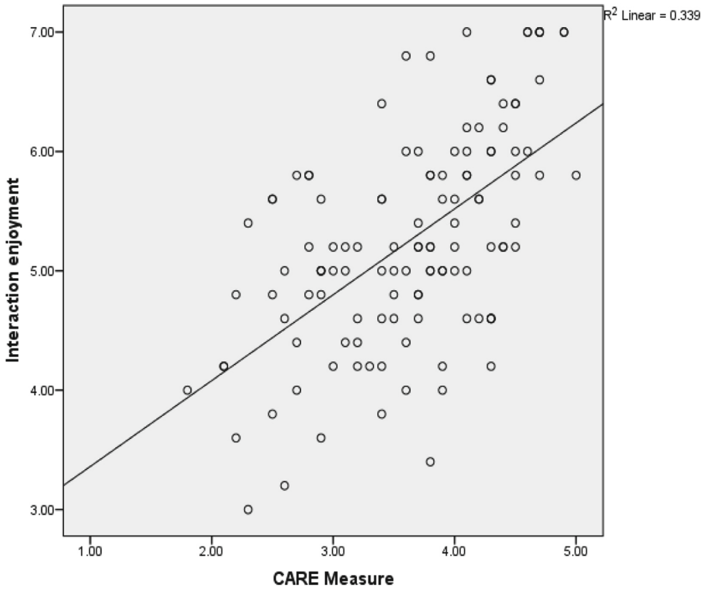

Fig. 1. Scatterplot with regression line for CARE Measure x Interaction enjoyment.

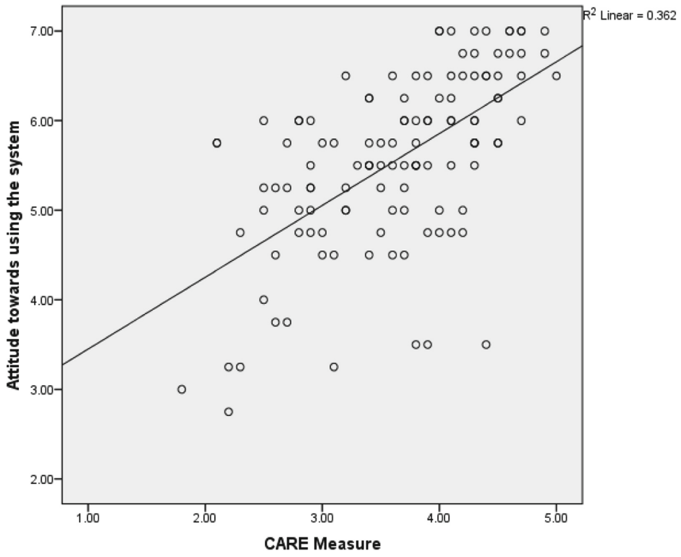

Fig. 2. Scatterplot with regression line for CARE Measure x Attitude towards using the system

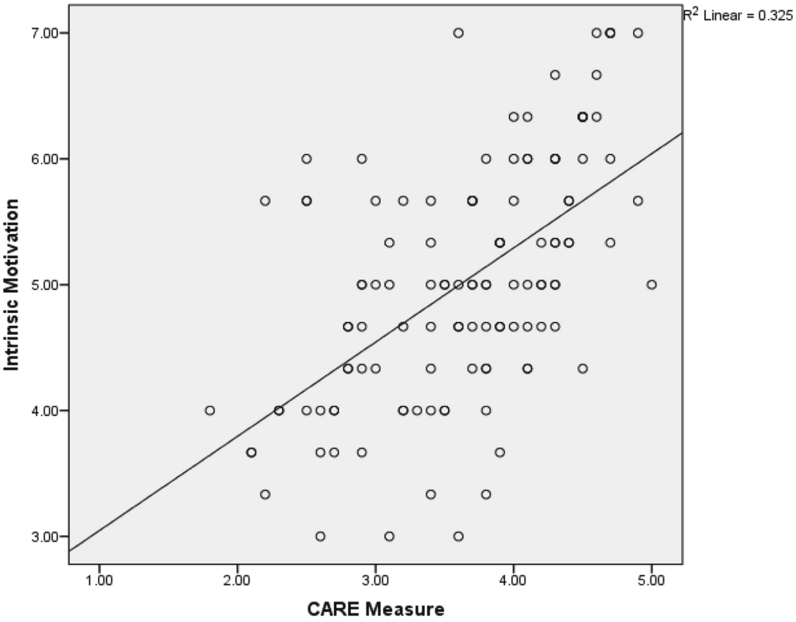

Fig. 3. Scatterplot with regression line for CARE Measure x Intrinsic Motivation.

3.6 Effect of Negative Attitude Towards Robots on Evaluation of the System

Since evaluation of interaction enjoyment, attitude, and intrinsic motivation did not differ across the experimental conditions, a sum score for each variable was created per participant and correlated with the ratings of the NARS questionnaire. Negative attitude towards automated systems correlated significantly negative for all three variables with the highest correlation found for interaction enjoyment ($r = -.508, p < .001$), followed by intrinsic motivation ($r = -.488, p < .001$) and attitude towards the system ($r = -.431, p < .001$). Participants with a negative attitude towards automated systems rated the VUI lower in interaction enjoyment, attitude towards using the system, and intrinsic motivation.

Correlation for the CARE measure were carried out separately for condition with and without empathic language cues, since the CARE measure differed significantly between both conditions as reported above. Attitude towards automated systems correlated significantly negative in both conditions, empathic $r = -.305, p = .018$, neutral $r = -.324, p = .012$. Participants who reported having a negative attitude towards automated systems perceived the systems as less empathic. These results indicate that participants general attitude towards automated systems influenced their evaluation of the VUI.

Order Effect

The scenarios used as well as manipulation of empathy was counterbalanced in the experiment. There was no difference across scenarios for CARE measure (t (60) = .28, p = .78), interaction enjoyment (t (60) = −1.29, p = .20), attitude (t (60) = 1.88, p = .06) and intrinsic motivation (t (60) =−.90, p = .37).

Session order had a significant effect on participants' evaluation of attitude towards the system (t (59) = 2.15, p = .035) and intrinsic motivation (t (59) = 3.119, p = .003). Participants had a more positive attitude towards the VUI in the first session (M = 5.67, SD = .901) compared to the second session (M = 5.46, SD = 1.05). Furthermore, participants reported a higher intrinsic motivation to use the VUI after the first session (M = 5.19, SD = .86) compared to the second session (M = 4.87, SD = .70). No effect was found for interaction enjoyment (t (59) = −.254, p = .800) and perception of empathy (t (59) = .965, p = 335.

4 Discussion

The goal of this study was to examine the effect of human likeness of voice and social behavior on the evaluation of interaction enjoyment, attitude and motivation to use a VUI. Special consideration was given to how the VUI was perceived by the users and how this perception affected users' evaluation.

Somewhat unexpectedly, human likeness of voice did not affect the evaluation of the VUI as proposed in Hypothesis 1. A machinelike voice received similar ratings regarding interaction enjoyment, attitude and motivation to use the VUI as a natural humanlike voice. The manipulation check showed that participants could recognize the voice as being humanlike or machinelike. This result contradicts previous findings [7, 15]. However, since voice was manipulated across groups, participants could not directly compare both types of voice with each other. If the voice is clearly understandable, participants do not seem to pay attention to the human likeness of the voice. The results might have differed if we would have chosen a within-design for voice manipulation as reported in other studies [7, 15]. This also raises the question of how experimental design affects the outcome of the study. This point is explained in more detail below. Since voice did not show any effect Hypothesis 4 and 5 were rejected as well.

Regarding the effect of social behavior, specifically expressing empathy, the results of this experiment are promising. The empathic expressions of showing interest, understanding the situation of the user, and offering help were reflected as expressions of empathy by the participants. The system that used empathic language cues was rated higher in empathy than the neutral system supporting Hypothesis 2. However, effects of empathy on other measures such as interaction enjoyment, attitude and motivation to use the system only appeared when looking at perceived empathy. Users' individual interpretation of a certain behavior as being empathic is the decisive factor for users' evaluation of the system as proposed in Hypothesis 3. These results suggest that individual differences must be factored in when incorporating social behavior into a system. It is not the behavior per se but rather its interpretation by the user that affect overall system evaluation. Whether a certain behavior is interpreted as an expression of empathy depends on the situational context, interpersonal relationship and role of the interaction partners [20–22]. These poses specific challenges for implementing social behavior into automated systems. Well defined areas where interaction follows a script, where user expectations are clear and social rules for accepted behavior are well defined seem appropriate. In addition, it is difficult to anticipate all eventualities that may occur during an interaction. For the user it is not only important what the system says, but also whether the system listens well to the requests and needs of the user.

Another result of our study was that the users rated the system more negatively in the second session compared to the first session regardless of the experimental condition. Anecdotal evidence suggests that users had little or no expectations in the first session. Many users were surprised that the system worked better than expected, which probably influenced the positive evaluation of the system in the first session. Due to their experience of the first session, users had higher expectations for the second session. In the second session, users became more aware of the limitations of the system, which could have led to a more negative evaluation of the system. In general, the role of user expectations is not yet well understood. It is unclear how these expectations change when interacting repeatedly with computer agents over time. Further research is needed in this area as most of these systems are intended for repeated use. The current study has shown that user expectations change and significantly affect the user's evaluation of a system regardless of the actual performance of the system.

In this regard it is also important to discuss the role of the experimental design on the results. Many studies using within-designs [13, 15, 19], where participants interact with various systems and then evaluate them. Through within-designs individual differences can be controlled and fewer participants are needed for a study. However, order effects must also be considered and should be reported. Exposure to one system affects user expectations, which then impacts the assessment of the next system. Changes in expectations can be avoided by a between-design. However, the role of individual differences must be considered as well. As the current study has shown, the attitudes of the participants to automated systems correlate with the evaluation of these systems. Participants with an initial negative attitude towards automated systems rated the VUI consistently more negatively than participants with a positive attitude towards automated systems. Researchers need to carefully consider the experimental design for their study, considering their key research objectives. However, in addition to the manipulated variables, researchers should also think critically about how the design affects the outcome of their study.

The current study used two specific scenarios (losing an item at the airport) to test the VUI. Since empathy is nested in a specific context, future studies should examine a variety of scenarios to control if this is a robust effect. An effect of voice was not found in this study probably because of the between-design. In the future we want to test if the results will be different if voice is manipulated as a within-factor. Furthermore, the machinelike voice was derived by altering the humanlike voice with a software. It is possible that the manipulation was not strong enough to show a difference. It was clear to participants that the humanlike voice was spoken by a computer agent. Creating a stronger contrast between humanlike and machinelike voice could produce different results and should be tested in further studies.

5 Conclusion

The study presented here examined the effect of human likeness of voice and social behavior (empathic) on users' evaluation of a VUI. The unique contribution of this study is that not only the effect of the behavioral manipulation of the system was studied, but also the importance of the perception of this behavior by users. More than behavioral manipulation, the individual perception and interpretation of this behavior influenced the assessment of the users. The study shows that individual differences play a decisive role in the evaluation of system behavior. Therefore, it is important to consider the interaction of human and computer agent as a dynamic process that is subject to constant change.

Whether empathy improves human-computer interaction may depend on the context of the situation. For computer agents offering a service, empathy can contribute to a positive evaluation of the interaction and might motivate the user to use the system again.

VUI must be able to adapt to changing request and needs of users. VUI should be sensitive to contextual variables and the specific interaction situation. Future research should focus more on the changing expectations of users, as these influence the motivation and attitudes of users towards such systems. Human-computer interaction is a dynamic process that changes in the course of interaction. Developers of VUI need to take this interaction dynamic into account when developing such systems.

Acknowledgements. We thank all the volunteers who participated in this study.

References

1. Reeves, B., Nass, C.I.: 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 (1996)
2. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)
3. Dzindolet, M.T., et al.: The role of trust in automation reliance. Int. J. Hum Comput Stud. **58**(6), 697–718 (2003)
4. Araujo, T.: Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput. Hum. Behav. **85**, 183–189 (2018)
5. Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. J. Exp. Psychol. Appl. **7**(3), 171–181 (2001)
6. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. J. Soc. Issues **56**(1), 81–103 (2000)
7. Louwerse, M.M., et al.: Social cues in animated conversational agents. Appl. Cogn. Psychol. **19**(6), 693–704 (2005)
8. Ruijten, P., Cuijpers, R.: Dynamic perceptions of human-likeness while interacting with a social robot. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 273–274. ACM, Vienna, Austria (2017)
9. von der Pütten, A.M., et al.: ‘It doesn’t matter what you are!’ explaining social effects of agents and avatars. Comput. Hum. Behav. **26**(6), 1641–1650 (2010)
10. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors **57**(3), 407–434 (2015)
11. Manzey, D., et al.: Human performance consequences of automated decision aids: the impact of degree of automation and system experience. J. Cogn. Eng. Decis. Making **6**(1), 57–87 (2012)

12. Mitchell, W.J., et al.: A mismatch in the human realism of face and voice produces an uncanny valley. i-Percept. **2**(1), 10–12 (2011)
13. Złotowski, J., et al.: Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. J. Behav. Robot. **7**(1), 55–66 (2016)
14. Stárková, T., et al.: Anthropomorphisms in multimedia learning: Attract attention but do not enhance learning? J. Comput. Assist. Learn. **35**(4), 555–568 (2019)
15. Ilves, M., Surakka, V.: Subjective responses to synthesised speech with lexical emotional content: The effect of the naturalness of the synthetic voice. Behav. Inf. Technol. **32**(2), 117–131 (2013)
16. Muresan, A., Pohl, H.: Chats with bots: balancing imitation and engagement. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6. ACM, Glasgow (2019)
17. Lottridge, D., et al.: Affective interaction: understanding, evaluating, and designing for human emotion. Rev. Hum. Factors Ergon. **7**, 197–237 (2011)
18. Leite, I., et al.: The influence of empathy in human–robot relations. Int. J. Hum Comput Stud. **71**(3), 250–260 (2013)
19. Niculescu, A., et al.: Making social robots more attractive: The effects of voice pitch, humor and empathy. Int. J. Social Robot. **5**(2), 171–191 (2013)
20. Butler, E.A.: Temporal interpersonal emotion systems: the “TIES” that form relationships. Pers. Soc. Psychol. Rev. **15**(4), 367–393 (2011)
21. Main, A., et al.: The interpersonal functions of empathy: a relational perspective. Emot. Rev. **9**(4), 358–366 (2017)
22. Butler, E.A.: Interpersonal affect dynamics: it takes two (and time) to tango. Emot. Rev. **7**(4), 336–341 (2015)
23. Kupetz, M.: Empathy display as interactinal achievements - multimodal and sequential aspects. J. Pragmatics **61**, 4–34 (2014)
24. Brave, S., et al.: Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int. J. Hum Comput Stud. **62**(2), 161–178 (2005)
25. Urakami, J., et al.: Users’ perception of empathic expressions by an advanced intelligent system. HAI, Kyoto (2019)
26. Syrdal, D.S., et al.: The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. Adaptive and Emergent Behaviour and Complex Systems (2009)
27. Mercer, S.W., et al.: The consultation and relational empathy (CARE) measure: Development and preliminary validation and reliability of an empathy-based consultation process measure. Fam. Pract. **21**(6), 699–705 (2004)
28. Sacco, D.F., Ismail, M.M.: Social belongingness satisfaction as a function of interaction medium: face-to-face interactions facilitate greater social belonging and interaction enjoyment compared to instant messaging. Comput. Hum. Behav. **36**, 359–364 (2014)
29. Davis, F.D., et al.: User acceptance of computer technology: a comparison of two theoretical models. Manag. Sci. **35**(8), 982–1003 (1989)
30. Davis, F.D., et al.: Extrinsic and intrinsic motivation to use computers in the workplace. J. Appl. Soc. Psychol. **22**(14), 1111–1132 (1992)
31. Damiano, L., et al.: Artificial empathy: an interdisciplinary investigation. Int. J. Soc. Robot. **7**(1), 3–5 (2015). https://doi.org/10.1007/s12369-014-0259-6